CLINICAL DECISION-MAKING AND PEDIATRIC BIPOLAR DISORDER

Melissa M. Jenkins

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology (Clinical Psychology).

Chapel Hill
2009

Approved by:

Eric A. Youngstrom, PhD

Enrique W. Neblett, Jr., PhD

Jen Kogos Youngstrom, PhD

ABSTRACT

MELISSA M. JENKINS: Clinical Decision-Making and Pediatric Bipolar Disorder
(Under the direction of Eric A. Youngstrom, PhD)

Clinical decision-making in mental health could greatly benefit from evidence-based decision tools, particularly in diagnosing challenging, high-stakes conditions such as pediatric bipolar disorder. The current literature indicates that clinicians are prone to a host of cognitive biases that impede optimal diagnostic and treatment decisions. These biases are especially salient in the assessment of bipolar illness. Bipolar disorder is frequently misdiagnosed, and recent evidence suggests that mental health professionals often overdiagnose bipolar in youths. Although actuarial approaches have taken root in the medical community to assess the likelihood of various conditions, the mental health field has not widely disseminated or implemented such strategies. In fact, little research has attempted to validate the clinical utility of actuarial assessment methods. This study examines the effectiveness of an actuarial approach in diagnosing pediatric bipolar disorder by comparing Bayesian estimates (i.e., actuarial approach) to the current gold standard in clinical assessment.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

FIGURE

LIST OF ABBREVIATIONS

BP-I	Bipolar I Disorder
BP-II	Bipolar II Disorder
BP-NOS	Bipolar Disorder Not Otherwise Specified
DSM	Diagnostic and Statistical Manual of Mental Disorders
KSADS	Schedule of Affective Disorders and Schizophrenia for Children
LEAD	Longitudinal Expert All Data
MH	Mental Health
PBD	Pediatric Bipolar Disorder

CLINICAL DECISION-MAKING AND PEDIATRIC BIPOLAR DISORDER

Decision-making algorithms can facilitate more timely, accurate, and effective decisions and can be applied to a number of situations, including inference, choice, group deliberations, and moral issues (Jenkins, Youngstrom, Youngstrom, & Algorta, 2008; Todd & Gigerenzer, 2007). Unfortunately, humans typically do not use purely rational or normative approaches in making decisions (Lau & Coiera, 2007; Tversky & Kahneman, 1974). This is evident and problematic in a variety of professional settings. For example, exorbitant fees result from unnecessary medical tests and procedures (Kraemer, 1992), and accurate and reliable assessment practices are recurring problems in mental health (MH) care (Youngstrom, 2007). The decision-making literature abounds with evidence-based (EB) decision strategies, such as Bayes' Theorem; however, the adoption of these strategies appears to get lost in translation. Rarely do these statistically savvy tools move from science laboratories into applied settings, such as community MH clinics, which could derive significant gain from their implementation. Current estimates indicate that approximately 14% of youths in the United States experience a moderate to serious MH problem (Kessler, Chiu, Demler, & Walters, 2005). Of those who seek treatment, research suggests a high rate of misdiagnosis at initial presentation (McClellan, Werry, & Ham, 1993; Reimherr & McClellan, 2004).

Clinical assessment is a major professional role largely unique to psychologists in the MH field (Krishnamurthy et al., 2004). Test interpretation is typically done using clinical judgment instead of actuarial approaches (Dawes, Faust, & Meehl, 1989; Meehl, 1954), which apply mathematical and statistical methods to assess risk. This finding is disconcerting for a number of reasons. First, there is poor agreement between clinical judgment and actuarial judgment (Dawes et al., 1989). Second, clinical judgment has been shown to suffer from cognitive biases and errors that can reduce accuracy in decision-making (Galanter & Patel, 2005; Youngstrom & Kogos Youngstrom, 2005). Third, research has shown that the validity of clinical judgment and the amount of clinical experience are unrelated (Lueger, 2002). In short, research indicates a tendency for clinical judgment to derail accurate decision-making, and this derailment does not self-correct with additional clinical experience.

Actuarial decision-making is most commonly used in the medical community to assess the likelihood of a particular disease. Specifically, actuarial methods take a Bayesian approach: risk of disease is quantified by combining base rate information with diagnostic test results, estimating the probability that a given patient has an illness (Hunink et al., 2001); (Sackett, Haynes, & Guyatt, 1991). Probabilities are then used to inform clinical decision-making. Recent studies indicate that actuarial approaches can be used in psychological assessment to improve assessment accuracy and increase agreement surrounding diagnostic decisions (Jenkins et al., 2008). These findings hold particular appeal for challenging, high-stakes diagnoses, such as pediatric bipolar disorder (PBD), which is associated with frequent misdiagnosis. A recent study found that for 52% of youth patients in a community MH setting, five or more years elapsed from the onset of symptoms before arriving at a bipolar

diagnosis (Marchand, Wirth, & Simon, 2006). Misdiagnosis and treatment delays can have harmful, if not life threatening, consequences for individuals with bipolar disorder, thus, intensifying the need for accurate and efficient decision-making for cases with bipolar.

Despite the potential for substantial improvement in assessment practices, to date almost no research has examined how actuarial decision-making performs in comparison to psychiatric assessments of real-world clients. More concretely, we do not know if actuarial risk estimates gel with diagnostic estimates generated from more sophisticated assessment methods, such as semi-structured approaches like the Kiddie-Schedule for Affective Disorders and Schizophrenia (KSADS) (Kaufman et al., 1997), above and beyond routine clinical judgment. Moreover, it remains unclear if different clinical presentations of psychiatric disorders affect the level of agreement between actuarial and clinical assessment methods. For example, there are narrow phenotypic versus more heterogeneous presentations of PBD (Leibenluft, Charney, Towbin, Bhangoo, & Pine, 2003); and it is uncertain if these related yet distinct clinical presentations influence or limit the precision of actuarial approaches. This consideration is highly relevant given that clinical assessment instruments have been shown to vary in sensitivity and specificity for different bipolar presentations. Specifically, in assessing the validity of the Mood Disorder Questionnaire (MDQ), Miller and others (2004) found that although the MDQ shows good sensitivity in bipolar I (BP-I) cases, it demonstrates less sensitivity in cases of other bipolar spectrum disorders such as bipolar II (BP-II), cyclothymic disorder, and bipolar disorder not otherwise specified (BP-NOS).

This project proposes to investigate two popular topics: PBD and evidence-based assessment. Specifically, this study will explore the effectiveness of a particular actuarial

approach, the nomogram, in estimating the risk of a complex, frequently misdiagnosed mental disorder, bipolar illness. The overarching goal of this project is to examine how the nomogram performs in comparison to the current gold standard clinical assessment of PBD. Specific aims include better understanding: (1) the agreement between PBD research diagnoses and actuarial risk estimates based on the combination of family history and test score on the Parent General Behavior Inventory (Youngstrom, Findling, Danielson, & Calabrese, 2001); (2) the agreement between "Longitudinal Expert evaluation of All Data" (LEAD) (Spitzer, 1983) confidence ratings and Bayesian estimates when an EB assessment intervention threshold model is applied; (3) the relationship between clinician confidence in bipolar diagnoses and the type of bipolar spectrum illness; (4) the role of bipolar type in the agreement between PBD research diagnoses and actuarial risk estimates; and, (5) the correlation between Bayesian risk estimates using independent, published diagnostic likelihood ratios and logistic regression estimates of the probability of having a bipolar diagnosis using optimal weights for the sample (generalizability).

Decision-Making in Mental Health

Decision-making is often complicated in MH care as a result of blurry psychiatric diagnostic criteria and the lack of irrefutable evidence for ruling psychiatric illness in or out (e.g., unreliability of symptom criteria; polytheticism). In contrast to distinguishing fatigue associated with iron deficiency from that of a thyroid condition, a simple blood test cannot be administered to differentiate "irritability" associated with PBD versus that of Attention Deficit Hyperactivity Disorder (ADHD), unipolar depression, ODD, or any one of several psychiatric disorders for which irritability is a major diagnostic symptom. Some authors propose that diagnostic interviews serve as proxies for laboratory tests (Zarin & Earls, 1993),

but many interviews have been shown to be unreliable (Piacentini et al., 1993). Assessing child psychopathology is often challenging and complex; consequently, clinicians routinely employ clinical judgment to navigate the decision-making process. Research shows, however, that clinical judgment is prone to a host of errors (Elstein & Schwartz, 2002), further complicating assessment. See Table 2 for examples of common faulty heuristics and cognitive errors in the decision-making literature.

MH professionals differ widely in their diagnoses of child psychopathology (Galanter & Patel, 2005). Inter-rater reliability among clinicians conducting unstructured interviews has been shown to be inadequate (Piacentini et al., 1993), and there is only moderate agreement between diagnoses based on standardized research interviews and those based on clinical chart reviews (Ezpeleta et al., 1997; Lewczyk, Garland, Hurlburt, Gearity, & Hough, 2003; Vitiello, Malone, Buschle, Delaney, & Behar, 1990; Vitiello & Stoff, 1997). Moreover, research shows that diagnoses and treatment vary across settings, which can lead to adverse outcomes for children (Galanter & Patel, 2005; Pappadopulos et al., 2002). One concrete example of this phenomenon involves prescription practices. Pappadopulos and colleagues (2002) found substantial variability in prescribed antipsychotic medications across different New York State inpatient long-term hospitals. This is problematic and potentially harmful because if clinicians evaluate the same individual and reach different diagnoses, one can infer that at least some individuals are misdiagnosed and therefore not receiving optimal services. In sum, evidence suggests that clinicians within and between clinical settings differ in their assessment approaches and interpretations of diagnostic information.

Further, many common assessment methods are not empirically supported (Fletcher, Francis, Morris, & Lyon, 2005; Neisworth & Bagnato, 2004); and attempts to change

clinician behavior often have not been successful (Galanter & Patel, 2005). Peterson (2004) comments, "For many of the most important inferences professional psychologists have to make, practitioners appear to be forever dependent on incorrigibly fallible interviews and unavoidably selective, reactive observations as primary sources of data" (p. 202). Some experts consider the current state of EB assessment as neglected (Mash & Hunsley, 2005). There is an apparent need to identify EB methods that appeal to clinicians and that they are willing to adopt into practice. This need is especially great for complex cases, such as those exhibiting bipolar symptomology.

Assessment of Pediatric Bipolar Disorder

PBD is a controversial diagnosis that has recently received a lot of attention in MH research and in the media (e.g., Kluger & Song, 2002; Papolos & Papolos, 1999). There has been a 40-fold increase in the number of young people diagnosed and treated for bipolar disorder in the span of a decade (Blader & Carlson, 2007; Moreno et al., 2007). The sizeable increase in rates of PBD is alarming. Two major concerns include the misdiagnosis of bipolar in youths (low accuracy, including low sensitivity) and the tendency for clinicians to overdiagnose PBD (low diagnostic specificity).

Misdiagnosed youths face a number of serious consequences, including delays in treatment and inappropriate treatments (Dunner, 2003). And, untreated cases may follow a progressive and deteriorating course of bipolar illness (Geller, Tillman, Craney, & Bolhofner, 2004). There is also some evidence that wrong medication, like antidepressants or stimulants, can possibly worsen outcome (cf. Joseph, Youngstrom, & Soares, 2009). Overdiagnosing or prematurely starting pharmacological treatment for bipolar disorder is dangerous because medications used to treat the illness can carry serious side effects (Wilens et al., 2003). There

is also evidence that suggests medications, such as antidepressants and/or methamphetamines, can bring about mania or mixed episodes in youth with bipolar (cf. Joseph, Youngstrom, & Soares, 2009). Suicidality represents another big concern associated with these medications (Olfson, Marcus, & Shaffer, 2006; Tondo, Isacsson, & Baldessarini, 2003). Overall, more research is needed to better understand the role of psychopharmacological agents in the treatment of PBD (Smarty & Findling, 2007). Weller, Danielyan, & Weller (2004) highlight the need for confident bipolar diagnoses before starting medication. Unfortunately, it is difficult to arrive at correct bipolar diagnoses with such confidence. For example, research indicates an average delay of more than 10 years between first episode and diagnosis of bipolar illness (Hirschfeld, Lewis, & Vornik, 2003). Another recent study found that 5 or more years elapsed from the onset of symptoms until making a bipolar diagnosis for 52.4% of youth patients in a community MH setting (Marchand et al., 2006).

Several key factors complicate the assessment and diagnosis of PBD. These factors include: (1) high rates of comorbidity, (2) overlap between PBD symptomology and symptomology of other conditions, (3) limitations of diagnostic tools, and (4) the cyclical nature of PBD illness (Bowring & Kovacs, 1992; Youngstrom, Findling, Youngstrom, & Calabrese, 2005). First, it is common for youths with PBD to also meet criteria for another psychiatric disorder (Findling et al., 2001; Kowatch, Youngstrom, Danielyan, & Findling, 2005) similar to what is found in adults (Kessler, 1999). High comorbidity is problematic because clinicians do not often witness a "typical" PBD presentation in isolation. This phenomenon of complex presentation coupled with the comparatively low prevalence of PBD can cause clinicians to only recognize the comorbid condition, neglecting PBD in their

case conceptualizations and treatment plans (Youngstrom et al., 2005). Second, overlapping symptomatology complicates diagnoses because it is hard to tease out bipolar symptoms from symptoms of more prevalent diagnoses, such as ADHD, unipolar depression, or conduct disorder. Third, common diagnostic instruments used in research, such as structured and semi-structured interviews, can be impractical for use in clinical settings due to issues of training, burden, and reimbursement (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006). Fourth, the cyclical aspect of PBD threatens the reliability of diagnostic impressions and can make diagnostic decisions difficult. For example, classic BP-I disorder can present as florid mania, severe depression, a mixed state, or as normal functioning (Kraepelin, 1921).

Gold Standard for Clinical Assessment

In light of not having clear-cut laboratory tests to validate psychiatric diagnoses or clinical assessment instruments, Spitzer (1983) introduced a provisional gold standard, the LEAD standard. LEAD encompasses three core concepts: "Longitudinal, Expert, and All Data" (p. 409). "Longitudinal" means that clients' symptoms are monitored over time. Past, present, and future symptoms are factored into diagnostic decisions (with diagnoses being revised in light of new information). "Expert" refers to clinicians who can make reliable diagnoses based on independent evaluation of the available data, comprehensive clinical interviews, and discussion with other experts around any diagnostic disagreement. Expert clinicians ultimately make consensus diagnoses that serve as the criterion measure. "All Data" refers to multiple sources of information, such as secondary informant reports from parents or teachers and data provided by other professionals (e.g., psychiatric history, etc.). In sum, LEAD standard represents a comprehensive and thorough approach to psychiatric

evaluation.

Several studies have since employed the LEAD standard (Klassen, Miller, & Fine, 2006; Miller, 2001, 2002; Miller, Dasher, Collins, Griffiths, & Brown, 2001; Peters & Andrews, 1995; Pilkonis, Heape, Ruddy, & Serrao, 1991). For example, Pilkonis and others (1991) investigated the reliability, stability, and clinical and predictive validity of Axis II diagnoses in patients with depression. Specifically, researchers tested the validity of two personality disorder instruments, the Personality Disorder Examination (Loranger, Susman, Oldham, & Russakoff, 1987), and the Personality Assessment Form (Pilkonis & Frank, 1988; Shea, Glass, Pilkonis, & Watkins, 1987) using the LEAD approach. Study findings provided preliminary support for using LEAD standard methodology. Further, Miller conducted a series of studies comparing the Structured Clinical Interview for the Diagnostic Statistical Manual for Mental Disorders-Clinical Version (SCID-CV) and computer-based decision support systems to consensus diagnoses (i.e., consensus diagnoses were arrived at using Spitzer's LEAD standard) (Miller, 2001, 2002; Miller, Dasher et al., 2001). For example, psychiatric diagnoses from a computer assisted diagnostic interview (CADI) were compared to consensus diagnoses from Spitzer's LEAD standard and were found to be in agreement = 86% and kappa = 0.81 ('excellent') (Miller et al., 2001). The Miller studies are highly relevant given that Miller used LEAD diagnoses to assess the validity of novel decision heuristics (i.e., a computer assisted diagnostic instrument), which represents a similar goal of the present proposal (i.e., compare LEAD diagnoses to Bayesian estimates).

Bipolar Phenotypes

In 2001, the National Institute of Mental Health (NIMH) Research Roundtable on prepubertal bipolar disorder agreed on two distinct phenotypes of bipolar disorder: *narrow*

and *broad*. The narrow phenotype consists of recurrent episodes of major depression and mania/hypomania (Nottelmann et al., 2001). According to this classification system, BP-I and BP-II fall under the narrow phenotype. The broad phenotype has various definitions; nevertheless, chronic mood instability or lability (versus discrete mood episodes) and irritability (minus euphoria or depression) represent common features of the broad phenotype. Diagnoses falling under the broad phenotype include BP-NOS, cyclothymic disorder, or subsyndromal bipolar disorder.

Despite the complexity involved in the assessment of bipolar disorder in children and adolescents, there is a growing consensus in the research community, at least, about the validity and clinical presentation of BP-I (Ghaemi et al., 2008; Youngstrom, Birmaher, & Findling, 2008). In particular, BP-I may be readily recognizable if symptom presentation closely aligns with the criteria outlined in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed. [DSM-IV]; American Psychiatric Association, 1994), including increased energy, feelings of grandiosity, and conventional symptoms of mania (excluding 'flight of ideas' and hypersexuality, which occurred much less frequently in study samples) (Axelson et al., 2006; Kowatch et al., 2005; Soutullo et al., 2005). Further, recent research indicates that MH professionals are more likely to agree on a bipolar diagnosis when the clinical presentation is classic BP-I (i.e., classical cases of mania) (Dubicka, Carlson, Vail, & Harrington, 2008).

The presence of family history of mood disorder, particularly family history of bipolar disorder, increases the likelihood of BP-I and appears to be recognized across different bipolar research camps as meaningful information in the diagnostic process (Youngstrom et al., 2008). At present, evidence suggests that a youth is roughly five times

more likely to have bipolar if he/she has a first degree (biological) relative with bipolar (Hodgins, Faucher, Zarac, & Ellenbogen, 2002; Smoller & Finn, 2003; Youngstrom & Duax, 2005). Overall, there is more agreement around narrow phenotype bipolar, especially in cases of BP-I, when family risk is present.

In contrast to the growing consensus around BP-I, the rest of the bipolar spectrum (i.e., BP-II and the "broad phenotype") continues to be more difficult to recognize and diagnose. Current diagnostic nosology is challenged by the heterogeneity of bipolar and the role of developmental influences on its clinical presentation (Leibenluft et al., 2003). Moreover, chronic and oscillating mood presentations associated with broad phenotype bipolar cause substantial impairment and seem to be more prevalent than BP-I (Lewinsohn, Klein, & Seeley, 1995; Youngstrom et al., 2008). Experts have various recommendations for addressing bipolar spectrum and the problems associated with current diagnostic definitions. For example, Leibenluft and others (2003) propose multiple phenotypes (i.e., narrow, intermediate, broad) based on characteristics of the manic or hypomanic episodes, while others argue that bipolar disorder represents a continuum of illness lacking any point of 'real cleavage' from unipolar depression (Phelps, Angst, Katzow, & Sadler, 2008). These different perspectives underscore the current state of the field; there is a dearth of empirical information to guide next steps in addressing diagnostic classification of broad phenotype bipolar (i.e., DSM-V). Despite the need for a greater understanding of this clinical population, clearly there is a group of youths who endure impairing bipolar symptoms but do not fit neatly into current diagnostic categories, no doubt complicating assessment, diagnoses, and treatment decisions. The lack of clear consensus and guidelines for training and assessment open the door for wide variations in clinical practice; and, there is evidence

of tremendous range of opinion among practicing clinicians about how to label cases with "soft spectrum" bipolar presentations (Dubicka et al., 2008; Jenkins et al., 2008).

Actuarial Decision-Making: The Nomogram

According to Sedlmeier & Gigerenzer (2001), "Unlike reading and writing, statistical literacy- the art of drawing reasonable inferences from such numbers, is rarely taught. The result of this has been termed 'innumeracy' (Paulos, 1988)." Statistical reasoning techniques could largely benefit decision-making in the clinical assessment of PBD and other challenging diagnoses. Bayes' Theorem is a highly accurate means of combining information about risk (Youngstrom et al., 2005). Evidence suggests that the presentation of information can significantly affect one's ability for learning how to use interpretative methods. For example, Bayesian computations appear easier to perform using natural frequencies than with probabilities (Sedlmeier & Gigerenzer, 2001). Children have successfully applied Bayes' Rule when information was presented in natural frequencies (Zhu & Gigerenzer, 2006). These findings support the feasibility of teaching statistically advanced techniques to individuals of varying educational, scientific, and occupational backgrounds.

Most training programs do not teach application of Bayes' Theorem, nor are these tools routinely used in spite of decades of research demonstrating their advantages in the laboratory (Gigerenzer, Hell, & Blank, 1988). Despite the evidence that people can be taught to think like "natural frequentists," a more user-friendly technique is needed for Bayes' Theorem to actually take root in real-world MH settings. Nomograms, which have gained popularity in EB medicine over the last ten years, are attractive tools because they incorporate Bayes' Theorem but require less mathematical computation- they function as a type of slide rule (Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000). In short, the

nomogram approach involves connecting the dots correctly to estimate revised, posterior probabilities, whereas computing natural frequencies involves first calculating the dots to connect (see Appendix A).

The nomogram is an EB decision aid that significantly increases assessment accuracy and significantly decreases interpretation variation among clinicians (Jenkins et al., 2008). Jaeschke, Guyatt, & Sackett (1994) recommend the "nomogram" as a simple, practical method for combining information about risk with the "diagnostic likelihood ratios" associated with test results or other clinical findings. The nomogram allows clinicians to work directly with probabilities without requiring any mathematical computation (Youngstrom & Duax, 2005). Although the nomogram approach is central in EB medicine for assisting with diagnostic and treatment decisions (Guyatt & Rennie, 2002; Straus, Richardson, Glasziou, & Haynes, 2005), it is not commonly used among psychologists and other MH professionals.

A Bayesian approach can be particularly effective in guiding challenging, high-stakes diagnoses, such as PBD. The nomogram correctly combines three pieces of information (i.e., base rate, familial risk, and test score) into consistent (less spread in opinion), unbiased (neither systematically over- or under-estimating risk), and efficient (using a parsimonious amount of information to arrive at the posterior probability) estimates (Jenkins et al., 2008; Youngstrom & Duax, 2005). The nomogram could also synthesize other sources of information. This estimate, the Bayesian posterior probability, can be used in the assessment of PBD to determine the likelihood that a youth has PBD, and to guide next steps in assessment and treatment. Given that PBD is a rare condition and that humans are more prone to commit faulty "cognitive heuristics" when the condition of interest is rare (Davidow

(Davidow & Levinson, 1993), the nomogram offers particular clinical utility in the assessment of PBD.

A recent clinical vignette study provides empirical evidence that the nomogram significantly increases assessment accuracy and decreases interpretation variation among national and international MH professionals (Jenkins et al., 2008). In addition, use of the nomogram generated positive feedback from a majority of professionals, many of whom were clinicians on the front line of service provision. Given that participants came from diverse clinical backgrounds, this finding suggests that the nomogram has the potential to be implemented and positively received in a variety of clinical settings.

More specifically, study findings were consistent with qualms that MH professionals are prone to overdiagnose PBD and to interpret identical information differently. When participants estimated the risk of PBD via clinical judgment alone, interpretations ran the full gamut (0-100). Giving professionals another piece of relevant information (i.e., a CBCL score) did not significantly improve diagnostic accuracy or increase consensus between groups. This finding is disconcerting in light of the dramatic increase in PBD diagnoses; it suggests that bipolar diagnoses made using clinical judgment risk inaccuracy and disagreement across MH professionals. However, the use of a nomogram led to dramatic improvements in interpretation of assessment information by practicing MH professionals. The nomogram resulted in increased accuracy and *consistency* (i.e., less variability in estimates around the "true" Bayesian probability). A brief (< 30 minute training) was sufficient to produce changes in accuracy and large effects in terms of self-reported learning. Despite the encouraging findings in the Jenkins et al. (2008) study which utilized a clinical vignette methodology, no research has investigated the role of Bayes' Theorem in assessing

real-world PBD cases. One important next step will be identifying cases that may not be well-suited for the tool (i.e., narrow versus broad bipolar phenotypes, etc.).

To assess the likelihood that a youth has PBD, clinicians can use the nomogram and arrive at a precise Bayesian estimate without doing any mathematical computation (Youngstrom & Duax, 2005). The nomogram combines the base rate, familial risk, and a standardized test score. To use the nomogram, a clinician would follow these steps:

1) Plot the base rate estimate on the far left line. This is the starting probability of PBD (i.e., the prevalence of PBD in a given clinical setting, geographic region, etc.). This estimate can be found in the literature or, a clinical setting can fairly easily compute a more tailored base rate given the number of PBD clients receiving services at their clinic. Base rates will fluctuate some depending on the setting. For example, clinical settings that offer more immediate and intensive care, such as inpatient settings, hospital emergency rooms, or bipolar specialty clinics, likely necessitate higher starting base rates because individuals seeking treatment at these types of service settings typically represent more severe populations and are thus at greater risk of having bipolar illness. Research has suggested base rates of 30% for inpatient settings (Carlson & Youngstrom, 2003), 34% for acute psychiatric hospitalizations (Blader & Carlson, 2007), and between 15 and 17% for specialty outpatient services (Biederman et al., 1996). It is typical that inpatient settings have higher base rates of PBD than outpatient settings. Youngstrom and Duax (2005) recommend a PBD base rate of 6% for outpatient settings, in the absence of any other information about the local setting.

2) Plot familial risk estimate on the middle line. Research indicates family history as the only risk factor adequately documented to justify integrating into clinical decision-making at present (Tsuchiya, Byrne, & Mortensen, 2003). EB familial risk estimates are as

follows: biological parent (or other first degree relative) with bipolar disorder = 5; second degree relative = 2.5; "fuzzy" family history of bipolar disorder = 2; no family history of bipolar disorder = 1 (Hodgins et al., 2002; Tsuchiya et al., 2003; Youngstrom & Duax, 2005). Note that second degree relatives are associated with a risk estimate half that of first degree relatives. By way of genetic composition, individuals' second degree relatives carry half the genes of first degree relatives; therefore, they contribute roughly half the risk of individuals' first degree relatives.

"Fuzzy" family history entails either a past diagnosis of uncertain validity, or else a different diagnosis for which bipolar is often mistaken in a minority population (e.g., schizophrenia or conduct disorder) (DelBello, Lopez-Larson, Soutullo, & Strakowski, 2001; Strakowski, McElroy, Keck, & West, 1996). See Methods for additional information regarding the "fuzzy" category.

3) Connect the left and middle dots and extend the line across the far right vertical axis. The number on the right axis provides the posterior probability or the Bayesian estimate of PBD. This estimate is fairly precise and will be close to the Bayesian exact estimate (Guyatt & Rennie, 2002). Any errors introduced will be due to visual interpolation between the anchors, and due to imprecision connecting the dots and extending the line.

If a clinician has the results of a standardized test score, such as a score from a parent report on the Child Behavior Checklist (CBCL) (Achenbach, 1991a), the Parent Young Mania Rating Scale (PYMRS) (Gracious, Youngstrom, Findling, & Calabrese, 2002), or the General Behavior Inventory (GBI) (Danielson, Youngstrom, Findling, & Calabrese, 2003; Depue et al., 1981) then this information can be included in the risk estimate as well. The above procedure can be repeated by using the posterior probability as the next starting point

for the left line and the diagnostic likelihood ratio associated with the test score for the middle line (Youngstrom & Kogos Youngstrom, 2005). Alternately, the DLRs could be multiplied together and the product used in single pass through of the nomogram; however, this approach would require the use of some arithmetic.

The risk estimates associated with family history and standardized test scores are diagnostic likelihood ratios (DLRs). A DLR is the sensitivity divided by the false alarm rate (i.e., complement of diagnostic specificity). The likelihood ratio indicates how often the risk factor, sign, or test score would occur in cases of bipolar disorder, compared to the rate of the same risk factor, sign, or score in cases without bipolar disorder (Jaeschke, Guyatt, & Sackett, 1994a, 1994b; Straus et al., 2005).

Threshold Model

The last step of the nomogram provides the Bayesian posterior probability, or the likelihood that a youth has bipolar disorder. This estimate falls somewhere between 0 (definitely does not have bipolar disorder) and 100% (definitely has bipolar disorder). Straus and others (2005) recommend using a threshold model to interpret this information (see Appendix B). This model consists of two thresholds, the test-wait threshold and the test-treatment threshold. These thresholds map on to three clinical activities (wait/assess/treat) and serve to guide next steps in clinical practice. A post-test probability that exceeds the test-treatment threshold informs the clinician to begin treatment, while a post-test probability that falls below the test-treatment threshold suggests the clinician assess further. In a similar vein, when the nomogram generates a probability below the test-wait threshold, this information informs the clinician to wait, or stop assessing for bipolar (i.e., the likelihood of bipolar illness is low enough that it does not make sense to continue assessing for bipolar). Estimates

that lie between the two thresholds (wait and treat) necessitate additional assessment because the likelihood of illness is not extreme enough to rule bipolar out (low end) or initiate treatment (high end).

A threshold model appeals for a number of reasons. First, this EB framework affords flexibility. The different thresholds can be tailored to important factors, such as client preferences. For example, given the serious side effects associated with the medications used to treat bipolar disorder, some clients and their families may want to delay psychopharmacological treatment (i.e., set a more conservative or higher treatment threshold), whereas others may be less concerned about potential side effects (e.g., weight gain and acne) and instead prefer to initiate psychopharmacological treatment immediately (i.e., set a more liberal or lower treatment threshold). Further, flexible thresholds may empower clients by giving them the opportunity to evaluate pros and cons of treatment (or further testing) and determine thresholds that meet their individual needs (Sackett et al., 2000).

Second, sophisticated calculations can be performed to establish test-treatment thresholds. Alternately, Straus et al. (2005) recognize and support intuitive test-treatment thresholds that rely on clinical expertise. Regardless of how test-treatment thresholds are decided, the threshold model represents an innovative approach that judiciously combines actuarial estimates with clinical insight. Practitioners in some clinical settings may want to implement an assessment intervention threshold model with wait/assess/treat thresholds based on statistical algorithms (see Gray, 2005 for more information about these calculations) and establish a policy of additional evaluation of individual cases in which clients evidence a likelihood of bipolar disorder at or near threshold cut-offs (i.e., incorporate

more clinical expertise for cases approaching test-wait and test-treat thresholds).

Third, using a threshold model to guide next steps in patient care helps prevent cognitive biases and faulty heuristics from creeping back into the decision-making process. For example, not providing wait/assess/treat ranges makes clinicians vulnerable to misinterpreting probabilities. Research shows that clinicians tend to interpret the same piece of diagnostic information very differently (Jenkins et al., 2008), so providing test-wait and test-treat thresholds can help guard against such inconsistencies in clinical practice. Moreover, if nomograms are adopted into practice for multiple disorders (e.g., ADHD and PBD) then it will be important for clinicians to recognize that post-test probabilities may have different consequences for different disorders (similar to how thresholds can be adjusted depending on patient preferences). For example, one might argue that a post-test probability of 85% for ADHD surpasses the test-treat threshold and triggers treatment, including psychostimulant medication, whereas an 85% post-test probability for bipolar may not be high enough to start treatment with mood stabilizers or atypical antipsychotic medications given the potentially serious consequences associated with prematurely starting medication. It is possible to incorporate information about the risks and benefits into calculations that select optimal thresholds (Kraemer, 1992; Straus et al., 2005), but a practical limitation is that normative data are often not available for quantitative estimates of risk and benefit. This does not prevent using the individual patient preferences, and skilled users of the nomogram could incorporate patient preferences into the decision-making process.

In sum, the threshold model highlighted by Straus and others (2005) has a number of strengths, including flexibility and clinical utility. Documenting probability estimates and thresholds can immediately transform decision-making into a more conscious and transparent

process (Youngstrom, 2008). Given the current state of assessment practices in MH, a user-friendly actuarial approach (the nomogram) used in conjunction with flexible interpretation guidelines (the threshold model) has significant appeal and could go a long way in improving MH services.

The Parent General Behavior Inventory (PGBI)

To achieve the most accurate actuarial estimate possible, it is important to select the best available screening instrument. In 2004, Youngstrom and others compared six index tests to determine diagnostic efficiency in predicting PBD, including the Parent Young Mania Rating Scale (PYMRS) (Gracious et al., 2002), the General Behavior Inventory (GBI) (Danielson et al., 2003; Depue et al., 1981), the Parent General Behavior Inventory (PGBI) (Youngstrom et al., 2001), Child Behavior Checklist (CBCL) (Achenbach, 1991a), the Youth Self Report (YSR) (Achenbach, 1991c), and the Teacher Report Form (TRF) (Achenbach, 1991b). This study has a number of noteworthy strengths. First, it used two relatively large samples and followed the STARD guidelines for reporting diagnostic test results (see Bossuyt et al., 2003). Second, this study involved the simultaneous inclusion of multiple index tests, which allowed for comparisons between different measures and different informants (parent, teacher, and youth). Third, evaluation of diagnostic efficiency used both global estimates and multilevel DLRs, which provide clinical utility for test interpretation. Finally, analyses were replicated on an independent sample of younger youth and bootstrapping procedures were incorporated providing nonparametric estimates of confidence intervals (i.e., less influenced by statistical outliers). Both of these features serve to enhance generalizability of study results.

Findings from the Youngstrom (2004) study have important clinical implications.

First, based on the areas under the curve (AUC) from Receiver Operating Characteristic (ROC) analyses and logistic regression results, parent report appears to be more useful than teacher or self report in detecting and diagnosing PBD in five to seventeen year olds. AUCs for the PGBI, CBCL, and PYMRS (.84, .78, and .80 respectively) were significantly larger than AUCs for the GBI, YSR, or TRF (.67, .71, and .70 respectively) in the older sample. Second, the PGBI represents the preferred rating scale in terms of changing the likelihood of PBD to a substantial degree when there are existing risk factors or clinical concerns of bipolar. The PYMRS has performed substantially less well in subsequent replication (Youngstrom et al., 2005). Third, none of the index tests alone are sufficient or intended as diagnostic instruments. Even though the PGBI demonstrated an AUC of .84, the most sensitive and specific of the six tests, this instrument used by itself is not sufficient to clinch a bipolar diagnosis.

In summary, the PGBI has evidenced impressive diagnostic efficiency in a well designed study comparing six popular index tests (Youngstrom et al., 2004). Parent report is currently the best source of information on questionnaires, and the PGBI appears to be a leading parent measure for detecting mania in PBD populations.

Limitations of the Nomogram

Despite the appeal of the nomogram as an effective, user-friendly decision aid, like all heuristics, it is not without limitations. First, the prevalence of PBD is a hot and contentious topic. In spite of the growing consensus about the validity of bipolar in children and adolescents, the research community continues to disagree about what constitutes broad phenotype bipolar and the frequency at which such diagnoses occur. This issue is problematic because: (a) accurate starting base rates necessitate familiarity with one's

clinical setting- specifically, knowing the current prevalence of PBD and being able to calculate the base rate; and, (b) base rates in any setting will most likely change as the field evolves, with changes in definitions and shifts in local practices or referral patterns. Periodically recalculating base rates will prove necessary to generating accurate Bayesian estimates with the nomogram.

A second limitation is that local base rates may be highly problematic in many clinics as a starting point because they will be based on clinical diagnoses. Therefore, local base rates are often likely to either underestimate (e.g., clinicians are conservative in diagnosing PBD) or overestimate (e.g., clinicians use broad definitions of bipolar) the actual base rate.

Third, the pieces of information used to arrive at the posterior probability may in some cases lack independence. This possibility represents a technical issue of the tool. For example, using the nomogram to combine responses from the same parent on multiple different questionnaires is likely inappropriate because the questionnaire scores will be highly correlated with each other, and will not contribute independent information. Logistic regression can accommodate correlated predictors and adjust the regression weights to consider the joint prediction and unique contributions of each predictor- something the nomogram cannot. DLRs are typically based on bivariate estimates of the sensitivity and specificity of predictors. This is most likely to be a problem when using multiple tests from the same informant (e.g., when a mother completes more than one questionnaire about the same child).

Significance and Broader Impact

Improving clinical decision-making has become a priority in the MH field (Chambers, 2008). Many assessment strategies that are efficacious in science laboratories are

not being regularly utilized in practice (Mass, 2003). The nomogram represents one of these underutilized strategies, but it is unique for a number of reasons. The nomogram is essentially free, requires brief training, and takes relatively little time to execute in practice. More importantly, it can help multiple stakeholders (i.e., therapists, clients, and clients' families) in the assessment of challenging diagnoses. Diagnostic decisions in cases with PBD represent one area where the nomogram could have a significant impact but this approach could also be applied to other often challenging diagnostic decisions, such as cases with ADHD. In addition, unlike other clinical decision-making tools, such as logistic regression or decision trees, which one must abandon when a predictor variable is missing, the nomogram is flexible and allows one to proceed with whatever pieces of information are available.

The nomogram has the potential to substantially improve the quality of MH services. However, before the nomogram can be successfully disseminated and implemented across clinical settings, it is necessary to first evaluate how the nomogram performs in comparison to the current reliable and valid assessment instruments. To our knowledge, no prior published research has compared the nomogram to empirically supported clinical assessment approaches. Gaining a better understanding of how actuarial approaches stack up to the gold standard (i.e., the KSADS) represents the building blocks for long-term improvement. In other words, the proposed research not only tests the nomogram's effectiveness; it seeks to investigate the conditions under which the nomogram is most useful (i.e., accurate) in guiding clinical decision-making. This understanding may in turn increase the rate at which actuarial approaches are understood, received, and adopted by MH professionals.

Hypotheses

1.1 Agreement between Clinician Confidence and the Nomogram.

There will be a correlation between the LEAD ratings (clinician confidence in bipolar spectrum diagnoses) and the Bayesian estimates (nomogram probabilities of having PBD) based on the combination of family history and test score on the Parent General Behavior Inventory.

1.2 Agreement about Next Clinical Action.

Applying an EB assessment intervention "threshold model" (wait/assess/treat) to LEAD confidence ratings and Bayesian estimates will evidence clinically significant agreement between the two assessment methodologies (e.g., Cicchetti et al., 2006).

1.3 Relationship between Clinician Confidence and Type of Bipolar.

LEAD confidence ratings for BP-I will be significantly higher than LEAD ratings for other bipolar spectrum diagnoses, including BP-II, cyclothymic disorder, and BP-NOS. Clinicians will be more confident in a bipolar diagnosis when the clinical presentation is classic BP-I (i.e., classical manic symptoms).

1.4 Potential Moderators of Agreement between Clinician Confidence and Nomogram.

Agreement between LEAD ratings and Bayesian estimates will be higher for BP-I than for broad phenotypes of pediatric bipolar disorder (i.e., type of bipolar will statistically moderate agreement). Significance of the interaction terms will be the direct test of the hypothesis.

1.5 Generalizability of Nomogram Approach.

Bayesian estimates using independent, published DLRs will correlate highly with logistic regression estimates of the probability of having a bipolar diagnosis using optimal weights for the sample (i.e., high degree of generalizability).

Methods

Procedure

Clinical Assessment.

All study procedures were approved by the Institutional Review Boards of University Hospital of Cleveland, Case Western Reserve University, and Applewood Centers, Incorporated. Written consent was obtained from parents or guardians for the participation of their children and written assent was obtained from all youths for their participation. The KSADS diagnostic interview (intake assessment) was administered by a highly trained research assistant to all participants and their families.

Parents completed a series of questionnaires while youths were interviewed, including the Parent General Behavior Inventory (PGBI) (Youngstrom et al., 2001), Child Behavior Checklist (CBCL) (Achenbach, 1991a), and the Parent Young Mania Rating Scale (PYMRS) (Gracious et al., 2002). Youth ages 11 to 17 completed the Youth Self Report (YSR) (Achenbach, 1991c) and the General Behavior Inventory (GBI) (Danielson et al., 2003; Depue et al., 1981), while the parent completed the KSADS interview. Rating scale responses were kept confidential so that youths and parents did not have knowledge of each other's responses. Upon completion of the interview, KSADS were scored. KSADS diagnoses were blind to rating scales.

After the client and his/her family completed the psychiatric evaluation, research diagnoses were reached through a Longitudinal Expert Evaluation of All Data (Spitzer, 1983) conference. All cases were reviewed by an expert consensus team, which always consisted of at least one licensed psychologist in addition to the rest of the members of the interview team for that given family. LEAD diagnoses were based on: (1) results from the KSADS; (2)

developmental history; (3) family history of mental illness; and, (4) psychiatric history, including any current diagnoses. Clinical chart reviews often provided information regarding youths' developmental and psychiatric histories. For each diagnosis, the expert clinician assigned a LEAD confidence rating based on how likely he/she viewed the diagnosis given all of the available information. These diagnoses and corresponding LEAD confidence ratings (likelihood of illness from 0-100%) represent the current gold standard in clinical assessment and will be compared to the Bayesian estimates.

Actuarial Assessment.

Procedures for actuarial assessments involved four steps. First, the prevalence or starting base rate for the study population was determined. Second, family history of bipolar illness and test scores on the PGBI were obtained for all cases in the aforementioned study. Third, family history of bipolar illness and test scores on the PGBI for these cases were translated into diagnostic likelihood ratios (DLRs). Fourth, the base rate of PBD and the DLRs for family history and PGBI scores were combined via Bayesian methods to provide the probability or likelihood of a bipolar diagnosis (i.e., the actuarial or Bayesian estimate). More detailed information about individual components of this process is provided below.

Prevalence.

Recent epidemiological studies indicate a lifetime prevalence of bipolar between roughly 6 and 11% (Judd & Akiskal, 2003; Merikangas et al., 2008; Soutullo et al., 2005). In outpatient clinical populations, evidence suggests prevalence estimates between 0.6 to 15%, depending on the diagnostic instrument, clinic specialization, and referral source (Geller, Craney et al., 2001; Lewinsohn et al., 1995; Strober et al., 1995). Because prevalence of bipolar disorder varies substantially by type of setting, it is important to consider the starting

base rate in light of clinical context. For example, the literature suggests a starting base rate of roughly 6% for outpatient settings (Geller et al., 2002; Youngstrom et al., 2005), 30% for inpatient settings (Carlson & Youngstrom, 2003), 34% for acute psychiatric hospitalizations (Blader & Carlson, 2007).

In the present study, the type of setting where participants were assessed was a blend of outpatient care and specialty clinics. Because these two settings differ substantially, there is no clear choice for the best starting base rate (i.e., there are no published data about recommended starting base rates for blended settings). We do know that a base rate of 6% is currently recommended for outpatient settings (Geller et al., 2002; Youngstrom et al., 2005) and that a base rate between 15 and 17% is recommended for specialty outpatient services (Biederman et al., 1996). Taking into account prevalence estimates found in the literature and base rate information for the present sample which was approximately 18%, a starting base rate of 12% was agreed upon for the present project. This process is an example of making an educated guess based on the literature and any specific information one can obtain from his/her clinical setting. Future studies should investigate base rates for blended settings as it represents an important consideration in using the nomogram or any actuarial assessment methodologies.

Diagnostic Likelihood Ratios (DLRs).

DLRs are used in actuarial assessment to help determine Bayesian estimates, or the probability that a youth has bipolar. The middle column of the nomogram contains the DLR information (Pepe, 2003). A DLR is the sensitivity divided by the false alarm rate. This can also be illustrated by dividing the percentage of cases with bipolar that would exceed the cutoff by the percentage of cases without bipolar that would also exceed the cutoff. Family

history and PGBI test scores can be translated into DLRs and plotted on the middle column of the nomogram (see Measures Administered).

Participants

A total of 643 youth, 5 to 17 years old, participated in the study. The mean age of study participants was 10.79 (*SD* = 3.46). Additional demographic characteristics of youth participants are provided in Table 2. The majority (90%) of youth were enrolled to participate in the present study by their mothers. Forty-two percent of participants' primary caretakers were single parents, 36% were married, 19% were divorced, and 4% were widowed. In regards to primary caretakers' highest level of education attained, 25% completed partial high school (grades 10 and 11) or less, 28% graduated from high school, 35% completed one to three years of college or trade school, 8% graduated from a four year university, 4% attended graduate school.

Half of youth participants' primary caretakers were currently unemployed. Twenty-four percent of primary caretakers had an estimated annual income under $5000. Twenty-nine percent earned $5000 to $14999, 20% earned $15000 to $29999, and the remaining 9% earned $30000 or more per year. Note that information pertaining to participants' financial status was only available on a subset of the sample (*n* = 167). More information regarding participants' socioeconomic characteristics is available upon request. See Table 3 for information pertaining to participant demographics and diagnostic characteristics.

Participants were recruited from two different clinical infrastructures. The first infrastructure was a community mental health center (CMHC) with four urban sites (Youngstrom, Kogos Youngstrom, & Starr, 2005). Invitations to participate in the study were extended to a random subsample of families presenting for outpatient treatment at the two

largest clinics. The second infrastructure was an outpatient academic medical center with more than a dozen different pharmacotherapy studies enrolling during the course of recruitment for the present study. Bipolar disorder (BP-I and cyclothymic disorder or BP-NOS), unipolar depression, ADHD, conduct disorder, and aggressive behavior regardless of diagnosis were target diagnoses for the treatment protocols. Presenting symptoms and willingness to participate in treatment protocols were the basis for recruitment. Families gained information about the different studies through different advertisements and referrals. Families interested in treatment studies partook in the diagnostic assessment for screening purposes and to provide baseline evaluations.

Referrals of children whose parents had a bipolar disorder and were participating in research or treatment at an associated adult mood disorders clinic enriched the academic sample. Additionally, under the auspices of a Child and Adolescent Psychiatric Clinical Research Center, youths and normal controls were recruited by word of mouth and study fliers to complete descriptive psychometric measures.

An outpatient clinic in an urban Midwestern city was the site for all assessments. Youths were included if they were between 5 years 0 months and 17 years 11 months of age, male or female, of any ethnicity, presenting for an outpatient assessment for which the youth gave written assent and his/her guardian gave written consent for participation, and both the youth and the primary caregiver were available for the assessment. Youth were excluded if they (or their respective caregivers) could not communicate orally at a conversational level in English to complete the interview, had a pervasive developmental disorder as evident by psychiatric history or psychiatric interview or having an Autism Screening Questionnaire score of 15 or higher (Berument, Rutter, Lord, Pickles, & Bailey, 1999), or there was

suspected moderate, severe, or profound mental retardation documented by educational history, standardized cognitive ability test scores of less than 70, or a Peabody Picture Vocabulary Test–Third Edition (Dunn & Dunn, 1997) score of less than 70. The same assessment procedures were administered to all eligible participants. This included the index tests and reference standard diagnostic interview and was regardless of participants' presenting symptoms or eligibility for treatment studies. Study design was "prospective" in that data collection was planned before assessment procedures were performed (Bossuyt et al., 2003).

Measures Administered

Reference Standard: Semistructured Diagnostic Interview Using the Schedule of Affective Disorders and Schizophrenia for Children. The Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime (KSADS-PL) (Kaufman et al., 1997) combined with the mood disorders module from the Washington University KSADS (WASH-U-KSADS) (Geller, Zimerman et al., 2001) was administered to all participants and their families. The WASH-U-KSADS consists of symptoms and associated features of depression and mania that other structured or semi-structured instruments do not capture. The KSADS is the most widely used semistructured diagnostic procedure for investigations of PBD (Nottelmann et al., 2001). BP-I, BP-II, cyclothymic disorder, and BP-NOS diagnoses were made in accordance with *DSM-IV* diagnostic criteria (American Psychiatric Association, 1994). The most frequent reason for diagnosing BP NOS was failure to meet strict DSM duration criteria (Leibenluft et al., 2003).

Training for research assistants consisted of having assistants observe five KSADS interviews (conducted by an experienced rater) and rate along. To graduate from training,

new raters were required to lead five KSADS interviews with experienced raters and to achieve an overall κ of > 0.85 at the symptom severity level and 1.0 agreement about the presence or absence of diagnoses. To maintain acceptable interrater reliability (κ > 0.85 about symptom severity) the research team held joint rating sessions. These sessions were scheduled monthly or at every 10th interview (whichever occurred second). One interviewer consistently worked with both the caregiver and youth to resolve discrepancies by employing clinical judgment as needed. In addition, per the joint reviews, average interrater agreement remained above a κ of 0.85 at the item level.

The Parent General Behavior Inventory (PGBI). The PGBI has been used to describe youths 5 to 17 years old with the primary focus of assessing the likelihood of a bipolar diagnosis based on high and low scores. Low scores decrease the odds that a given youth has a bipolar disorder, while high scores increase the odds. See Table 4 for DLRs associated with PGBI scores (see Youngstrom et al., 2004 for details).

Mini International Neuropsychiatric Interview (MINI). The MINI is a brief standardized diagnostic interview with documented reliability and validity (Sheehan et al., 1998) that was developed in 1990 by psychiatrists and clinicians for DSM-IV and International Classification of Diseases (ICD-10) psychiatric disorders. The administration time for the MINI is approximately 20 minutes. It is considered a valid and more time-efficient instrument than the Structured Clinical Interview for DSM diagnoses (SCID-P) and the Composite International Diagnostic Interview (CIDI). The MINI has been validated against the SCID-P in English, French, and Japanese (Lecrubier et al., 1997; Otsubo et al., 2005; Sheehan et al., 1997) and against the CIDI in English, French, and Arabic (Kadri et al., 2005; Lecrubier et al., 1997).

Family history of bipolar disorder was based on information obtained from the MINI which was administered by a trained rater. The MINI was first administered to the person who brought the youth in for evaluation and then this person was asked to answer similar questions about as many other family members as possible using a modified Family History – Research Diagnostic (Andreasen, Endicott, Spitzer, & Winokur, 1977). It should be noted that history of bipolar disorder may be incomplete; most of the time, raters only collected family history information for youths' relatives currently residing in the family household.

Fuzzy family history of bipolar disorder entails either a past diagnosis of uncertain validity, or a different diagnosis for which bipolar is often mistaken in an ethnic minority population (Youngstrom & Duax, 2005). The fuzzy category denotes an intermediate change in the DLR of bipolar in the proband: the risk (and the associated DLR) are lower than if the relative had a clearly documented case of bipolar disorder, but this risk is still not neutral and is higher than if the family had no history of mental health issues. It was not possible to identify the race/ethnicity of participants' relatives in the present study who received a clinical diagnosis on the MINI (see Limitations).

Of significance, the literature indicates that African-American populations are at high risk of being misdiagnosed with schizophrenia instead of a correct diagnosis of BP (Mukherjee, Shukla, Woodle, Rosen, & Olarte, 1983; Neighbors, Trierweiler, Ford, & Muroff, 2003; Neighbors et al., 1999). African-Americans are also less likely than Caucasian patients to receive bipolar diagnoses when all subjects present with psychotic mania (Strakowski et al., 2003; Strakowski et al., 1996). Moreover, there is evidence that Afro-Caribbean patients with mania are diagnosed with schizophrenia at high rates (Kirov & Murray, 1999). In light of these findings and the fact that youth participants in the current

study sample are predominantly African-American, any first degree relative with a diagnosis of any type of schizophrenia on the MINI was classified as "fuzzy" and assigned a DLR of 2.5.

A few additional disorders that qualified first degree relatives for fuzzy status warrant discussion. These disorders include: mood disorder NOS; mood disorder with psychotic features; brief psychotic disorder; psychotic disorder NOS; depression with atypical features; and, delusional disorder NOS. These disorders were included in the fuzzy category for two main reasons. First, differential diagnosis for this group of disorders is often challenging (Anthony et al., 1985; Helzer et al., 1985); in fact, there is evidence that clinicians are prone to misdiagnose these disorders for a variety of reasons (Ghaemi, Sachs, & Goodwin, 2000; Sprock, 1988). Given that differential diagnosis for these disorders would likely include bipolar disorders (i.e., manic symptoms frequently overlap with other psychotic disorder symptoms), it seems prudent to include these disorders in the fuzzy criterion for the purposes of the given project.

The MINI was first administered to the person who brought the youth in for evaluation and then this person answered similar questions about other family members. The precision of the family history information about other family members may be diluted as a result of retrospective report coming from a third party, which has been shown to be problematic (Andreasen et al., 1977; Weissman et al., 2000; Weissman et al., 1987). Because of the potential uncertain validity of diagnoses for "other" family members, including bipolar diagnoses that do not fit classic bipolar (e.g., BP-II), it seems most appropriate to assign broad phenotype bipolar diagnoses as well as other disorders that can resemble BP-I (e.g., mood disorder with atypical features) to fuzzy status.

Of significance, only first degree relatives were eligible for "fuzzy" status. Assigning first degree relatives with the aforementioned disorders to fuzzy status provides diagnostically useful information (i.e., a change in the diagnostic odds by 2); however, taking a similar approach for second degree relatives does not provide statistically or clinically meaningful information. More specifically, fuzzy status for second degree relatives would be associated with a DLR of 1, which would have a neutral influence on the actuarial estimate (i.e., would not increase or decrease the likelihood that a youth has PBD).

Overall, classification of family risk status for this project took a largely conservative approach. For example, only first degree relatives with BP-I were associated with a 5x increase in odds, whereas first degree relatives with other diagnoses on the BP spectrum were treated as "fuzzy" (a DLR of 2). Further, only first degree relatives with frequently misdiagnosed mental disorders (e.g., schizophrenia) were assigned to fuzzy status, whereas second degree relatives with similar disorders were not (i.e., aunts or uncles with schizophrenia were not included in analyses). See Table 5 for frequencies of EB familial risk estimates for the present study sample, and see Table 6 for DLRs associated with MINI diagnoses for first and second degree relatives.

Results

Power Analysis

There was more than adequate power to detect effects for all primary analyses (Donner, 1998; Faul, Erdfelder, Lang, & Buchner, 2007). A sensitivity analysis was run for Pearson's correlation, analysis of variance (ANOVA), and regression (Hypotheses 1, 3-5), and Donner (1998) was referenced to determine power for the analysis which used Cohen's kappa coefficient (Hypothesis 2).

First, for the hypothesis that used Pearson's correlation to examine agreement between clinician confidence and the nomogram, the power sensitivity analysis indicated that we had 80% power to detect a small effect size (r = .11) (Cohen, 1988) for the given sample size (N = 643) and alpha (.05, 2-tailed).

Second, according to Donner (1998), power was adequate for the hypothesis that used Cohen's kappa coefficient to examine agreement between clinician confidence and the nomogram in regards to next clinical action. Sample sizes of 203, 96, and 83 are 80% powered to detect effects of .1, .3, and .5 respectively (for alpha .05, two-sided) (Donner, 1998). Our sample size exceeded 203, suggesting we had more than 80% power to detect a small effect size.

Third, a planned comparison tested whether clinician confidence for BP-I was significantly higher than clinician confidence for other types of bipolar. The sensitivity analysis for ANOVA indicates that we had 80% power to detect a small to medium effect size (r = .23) (Cohen, 1988) for the given sample size (n = 149) and alpha (.05, 2-tailed). The sample is substantially smaller than the total sample because it is a subgroup of individuals with bipolar. Post-hoc multiple comparisons of clinician confidence by type of bipolar were

more than 95% powered according to a post hoc power analysis using the observed effect sizes.

Fourth, OLS regression was used to examine potential moderators of agreement between the nomogram and clinician confidence. We had 80% power to detect a small effect size ($f^2 = .06$) (Cohen, 1988) for the given sample size ($n = 136$) and alpha (.05, 2-tailed). This sample ($n = 136$) is a subgroup of individuals with a bipolar diagnosis so it is smaller than the total sample.

Lastly, Pearson's correlation quantified the generalizability of the nomogram approach. Results from this sensitivity analysis indicate that we had 90% power to detect a small effect size ($r = .13$) (Cohen, 1988) for the given sample size ($n = 624$) and alpha (.05, 2-tailed).

Descriptives and Missing Data

Analyses began with descriptive statistics. After acquiring means and standard deviations, skewness and kurtosis were used to examine the normality of distribution and the extent to which distributions differed substantially from approximately normal distributions. Minimum and maximum values were obtained for each variable and checked to make sure that they represented possible values. Impossible values were evaluated for possible data collection or data entry errors.

Family history information and research diagnoses (see Table 2 and Table 4) were available for all participants. In terms of descriptives and missing data for quantitative variables, three variables were of interest: LEAD ratings, Bayesian estimates, and PGBI test scores. None of these variables was normally distributed. See Table 7.

As expected, the distribution of clinician confidence ratings was positively skewed

moderately peaked distribution. The majority of the sample (80%) did not have bipolar, and therefore was assigned a confidence rating of 0. This means that most of the sample received a low score, making the median less than the mean, causing a positively skewed distribution.

Likewise, the distribution of nomogram estimates was positively skewed; however, not to the same degree as clinician confidence ratings. The majority of the sample did not have a family history of bipolar illness (i.e., not increasing the odds of a youth having bipolar) and the average PGBI score was 21.28, or within the neutral range (i.e., not significantly increasing or decreasing a youths' odds of a bipolar diagnosis). Because 12% was used as the starting base rate for bipolar disorder, and we know that the majority of nomogram estimates were not significantly influenced by the DLRs associated with family history and PGBI scores, it makes sense that the median for this variable was 13.20. Moreover, the minimum nomogram estimate was never 0, which likely prevented the distribution from appearing as skewed as the distribution for clinician confidence ratings (for which it was possible for diagnoses to receive a confidence rating of 0). Kurtosis was negligible for nomogram estimates.

The distribution for PGBI scores is also positively skewed, indicating that there were a few extremely high scores on the PGBI. Taking into account the prevalence of bipolar disorder in the present study (18%) and the impressive discriminative validity that the PGBI has evidenced in past studies (see Youngstrom et al., 2004), it is not surprising that overall more youths scored in the neutral range (21-30 for 5-10 year olds and 16-24 for 11-17 year olds) or lower (i.e., scores which suggest a lower probability of having bipolar). In other words, if only 18% of the sample was positive for a bipolar diagnosis, then we would predict that on a "good" measure (such as the P-GBI) most participants' scores would not surpass the

neutral threshold. In other words, the "base rate" and test "level" should be similar on a good test (see Kraemer, 1992). Kurtosis was negligible.

Less than 2% of data were missing. The rate of missing data points was calculated by multiplying the number of variables by the number of participants and then dividing by the number of complete data points. Given that the amount of missing data was small, missing data were excluded listwise. This approach provides less bias than pairwise deletion and is adequately suited for small amounts of missing data (Allison, 2002). Missing data did not impede analyses.

Agreement between Clinician Confidence and the Nomogram

Pearson's correlation tested the prediction that there was an association between the LEAD rating and the Bayesian risk estimate of having PBD based on the combination of family history and test score on the PGBI. The LEAD ratings correlated $r = .37$ with the Bayesian estimates, indicating a medium positive relationship (Cohen, 1988). Figure 1 superimposes the test-wait and test-treat thresholds on the scatter plot displaying Bayesian estimates and LEAD estimates. The Bayesian ratings were generally more conservative, almost never crossing the treatment threshold, and often falling in the low or indeterminate ranges for cases where the LEAD confidence ratings were high. This pattern suggests: (a) the Bayesian approach by itself would not be sufficient to initiate treatment, at least not with the current set of inputs; (b) additional information captured in the LEAD process may be necessary to cross the treatment threshold; and, (c) the Bayesian approach does not appear likely to generate "false positives" that would be exposed to unwarranted treatments.

Agreement about Next Clinical Action

Cohen's kappa coefficient was used to test if applying an EB assessment intervention

"threshold model" (wait/assess/treat) to LEAD confidence ratings and Bayesian estimates showed clinically significant agreement between the two assessment methodologies. Results indicate a $\kappa = .21$ when test-wait threshold and test-treat threshold were set at 25 and 85, respectively. Using benchmarks described in Cicchetti's recent review of levels of agreement (Cicchetti et al., 2006), LEAD confidence ratings and Bayesian estimates evidence fair agreement when a threshold model approach was taken. It should be noted that $\kappa = .21$ is on the cusp of slight and fair agreement (Landis & Koch, 1977) and that by other standards of clinical significance $\kappa < .40$ is considered poor (Cicchetti & Sparrow, 1981; Fleiss, 1981, 2003). Table 8 provides a detailed comparison of the clinical decisions generated by each approach.

Relationship between Clinical Confidence and Type of Bipolar

ANOVA compared mean confidence ratings between the BP-I group ($n = 28$), BP-II group ($n = 17$), cyclothymic disorder group ($n = 50$), and BP NOS group ($n = 54$). Confidence ratings between groups were significantly different, $F(3,145) = 7.72$, $p < .0005$. Planned comparisons examined the relationship between confidence ratings for BP-I and the other bipolar disorders. Results indicate that confidence ratings for BP-I were significantly higher than confidence ratings for cyclothymic disorder, $t(41.92) = -2.12$, $p < .05$, and BP NOS $t(57.22) = -3.56$, $p < .005$, but not for BP-II, $t(33.40) = -.82$, $p > .05$.

Because Hypothesis 3 was very focused, predicting significantly higher LEAD confidence ratings for BP-I versus other types of bipolar disorders, planned comparisons compared means (i.e., planned tests were determined before looking at the data). Nevertheless, the statistically significant effects in ANOVA coupled with the exploratory

nature of this project encouraged the research team to examine other relationships between clinician confidence and type of bipolar using post hoc analyses to make multiple comparisons across bipolar groups. Because of significant findings on Levene's test of equality of variances, $p < .005$, the Games-Howell statistic, which does not assume equality of population variances, compared means. Results from Games-Howell indicate that confidence ratings for BP NOS were significantly lower than confidence ratings for BP-II, $p < .05$, and cyclothymic disorder, $p < .05$. Figure 2 shows "box and whisker" plots for both Bayesian and LEAD confidence ratings by diagnosis.

Potential Moderators of Agreement between Nomogram and Clinical Confidence

An OLS regression approach tested the prediction that type of bipolar moderates agreement between LEAD confidence ratings and Bayesian risk estimates. Specifically, the Bayesian estimate predicted the LEAD confidence score, along with dummy codes for bipolar type, and interaction terms for bipolar type with Bayesian estimates. The significance of the interaction terms was the direct test of the hypothesis. None of the interaction terms were significant, indicating that type of bipolar did not statistically moderate agreement between the LEAD and Bayesian estimates. The actuarial estimate was also not a significant predictor; however, this can be explained by its redundancy with the dummy codes for diagnoses. See Table 9 for unstandardized regression weights and associated tests of significance. Figure 3 shows the regression lines for the interaction terms: diagnosis by actuarial Bayesian estimate.

Generalizability of Nomogram Approach

Pearson's correlation tested if Bayesian risk estimates using independent, published DLRs correlated with logistic regression estimates of the probability of having a bipolar

diagnosis using optimal weights for the sample (i.e., generalizability). The two estimates were highly correlated, as predicted $r = .82$, indicating a strong positive relationship (Cohen, 1988). Figure 4 shows the association between predictions based on external benchmarks versus regression weights optimized for the present sample.

Discussion

Current Literature

PBD and EB assessment have been in the spotlight of both clinical and research communities as well as the media and popular press (e.g., Kluger & Song, 2002; Papolos & Papolos, 1999). A recurring theme emerges: clinical decision-making could greatly benefit from EB decision tools, particularly in diagnosing difficult, high-stakes conditions such as PBD. To the best of the author's knowledge, to date no research has examined the relationship between bipolar diagnoses using the current gold standard for clinical assessment (a diagnosis based on the LEAD standard) versus taking an actuarial approach. The difficulty diagnosing PBD is well documented (Ghaemi, Sachs, Chiou, Pandurangi, & Goodwin, 1999; Hirschfeld, Calabrese et al., 2003); however, there is a dearth of information regarding the utility of EBA decision aids, such as the nomogram.

Despite large gaps in the extant literature, a few key findings warrant attention. Research suggests that clinicians are vulnerable to cognitive biases that often result in suboptimal diagnostic decisions (Croskerry, 2002; Galanter & Patel, 2005). PBD diagnoses have risen at alarming rates, with some estimates showing as much as a 40-fold increase in diagnoses over the last decade (Blader & Carlson, 2007; Moreno et al., 2007). Evidence indicates that PBD is frequently misdiagnosed (low accuracy, including low sensitivity), and research suggests that practitioners often overdiagnose PBD (low diagnostic specificity).

Present Study

The overarching goal of the present study was to compare the current gold standard for clinical assessment of PBD (a LEAD diagnosis based on a KSADS interview with collateral information and treatment history) to an innovative actuarial approach. The first

three study hypotheses tested particular aspects of the relationship between clinical and actuarial assessment methodologies, including agreement between clinician confidence and the nomogram; agreement about next clinical action; and, the relationship between clinician confidence and the nomogram. The final two study hypotheses examined relevant aspects of the individual approaches. Specifically, these hypotheses investigated potential moderators of agreement between the nomogram and clinician confidence, and the generalizability of the nomogram approach. Study findings have important research and clinical implications which are highlighted below.

Study Findings

A medium positive correlation was found between LEAD confidence ratings and Bayesian risk estimates. This finding supports the study prediction that clinician confidence and the nomogram would agree, suggesting a relationship between potent risk factors and clinician confidence. It is important to highlight, however, that clinician confidence was informed by the LEAD process- clinicians' ratings reflect findings from the KSADS interview, detailed family history, and clinical chart information. In short, clinician confidence reflected all available data per the LEAD standard and not just risk factors such as family history (also note: clinician confidence is also not equivalent to clinical judgment in the usual sense- see Meehl, 1954). The relationship between clinical and actuarial estimates would likely differ if clinicians did not have this additional information (e.g., KSADS findings, etc.) (see Jenkins et al., 2008). For example, one might argue that the correlation between clinical and actuarial estimates would be weaker if clinician confidence was blind to KSADS results and information from the clinical chart review as well as the case conference (see Spitzer, 1983). On the contrary, one might hypothesize that the relationship between the

two would be stronger. In explanation, given that LEAD ratings tended to be higher than Bayesian predictions, clinicians may have been less confident in their bipolar diagnoses without the additional "supporting" information from the KSADS (i.e., LEAD ratings might have been lower overall and more closely aligned with actuarial estimates which tended to be more conservative).

When a threshold model was applied to LEAD confidence ratings and Bayesian estimates (Hypothesis 2), these different approaches evidenced fair agreement (Cicchetti et al., 2006; Landis & Koch, 1977). In other words, when estimates from both approaches were translated into clinical activities (e.g., a score of 85 is associated with beginning treatment for PBD), clinical and actuarial methodologies agreed to a moderate extent. This finding has important clinical implications. Gold standard clinical assessments are often not feasible in real-world practice. For example, issues related to insurance reimbursement and staff training can hinder a comprehensive assessment such as the KSADS. Moreover, many clinical settings lack the resources to ensure acceptable administration of semi-structured diagnostic interviews which can result in unreliable diagnostic impressions and inappropriate treatment plans. The finding that the clinical and actuarial approaches agree at all both in terms of overall concordance and also in terms of deciding next clinical activities suggests that the nomogram represents a promising decision aid in settings that cannot routinely employ gold standard assessments. Because agreement is only found to be fair, clearly more work is needed to understand differences between these two approaches and what factors might further increase agreement.

BP-I confidence ratings were significantly higher than confidence ratings for cyclothymic disorder and BP NOS but not for BP-II. This finding suggests that expert

clinicians may have greater confidence diagnosing more "classic" presentation of bipolar illness. Moreover, confidence ratings for BP NOS were significantly lower compared to BP-I, BP-II, and cyclothymic disorder. This finding indicates that clinicians are less confident when diagnosing bipolar illness that does not fit into current categorical definitions of the disorder. In other words, bipolar presentations that deviate from DSM criteria may be associated with less diagnostic confidence.

Contrary to study predictions, the type of bipolar did not significantly moderate the level of agreement between LEAD confidence ratings and Bayesian estimates (i.e., the interaction terms were not significant). This finding is surprising given that LEAD confidence ratings varied by the type of bipolar (e.g., BP-I confidence ratings were significantly higher than BP NOS confidence ratings). Apparently, type of bipolar may affect clinicians' diagnostic confidence but it does not influence the degree to which actuarial and clinical assessment methodologies agree. This finding suggests that the nomogram approach is not limited to BP-I presentations; it can potentially be used in the assessment of any bipolar spectrum disorders. Clearly, more research is needed to support these preliminary findings; nevertheless, it is encouraging that a decision aid exists that may help clinicians better assess "soft spectrum" bipolar presentations that continue to be very difficult to recognize and diagnose.

As hypothesized, Bayesian risk estimates were highly correlated with logistic regression estimates. Theoretically, regression yields more precise predictions than the nomogram. In lay terms, regression creates a customized equation that best fits a given sample in order to most accurately predict bipolar status. The nomogram generates estimates using published DLRs, drawn from an independent sample, rather than using weights

developed and optimized to the new sample. Although regression can accommodate correlated predictors and adjust the regression weights to consider the joint prediction and unique contributions of each predictor- something the nomogram cannot do- it has noteworthy limitations.

In particular, if any predictor is missing, then regression loses its clinical utility (i.e., statistically, the approach must be abandoned). For example, if family history information is not available, one cannot use a regression model built with family history included as a predictor because all of the other regression weights in the model were estimated contingent on controlling for family history. The nomogram, however, is more flexible and allows one to proceed with the information available. Also, given that professionals have historically struggled with incorporating statistical approaches in their decision-making (Dawes et al., 1989; Gaissmaier & Gigerenzer, 2008), it is unlikely that clinicians would use regression in the assessment of PBD- another unattractive feature of a regression approach.

The finding that regression, a sample-specific approach that is very precise but relatively inflexible, and Bayes' Theorem, a non-sample specific approach that is very flexible, produced highly correlated estimates suggests that the nomogram has a high degree of generalizability. In other words, the nomograms estimates can be applied in new samples and yet perform at an acceptable level, showing a strong positive relationship with regression estimates.

Study findings cannot be discussed without considering the present study sample which was predominantly African-American. Given the well-documented difficulties assessing bipolar using other methods in minority samples (especially in the case of African-Americans), it is all the more impressive how well the nomogram performed. Actuarial

assessment methods may help decrease the rate at which clinicians misdiagnose African-Americans and other minority populations by supplanting faulty clinical judgment with more EB strategies. For example, if an individual presents with a family history of mental illness and/or receives a positive test score on a bipolar screening instrument, the nomogram generates the same probability of bipolar disorder regardless of race/ethnicity. Further, the Bayesian approach has outperformed clinical judgment in a previous vignette study involving an African-American youth (Jenkins et al., 2008).

Limitations

The present findings need to be interpreted in light of the study's potential limitations. First, study results may in fact demonstrate weakened effect sizes as a result of taking a conservative approach in defining family history of bipolar illness. For example, only participants with first degree biological relatives with BP-I were assigned a DLR of 5, and not first degree biological relatives with another BP diagnosis (i.e., BP-II, cyclothymic disorder, or BP NOS). In turn, first degree relatives with BP-II and broad phenotype bipolar were assigned to the fuzzy category which is associated with a substantially smaller DLR of 2. Given that LEAD confidence ratings were in general higher than actuarial estimates, it is plausible that only assigning first degree relatives with BP-I to DLRs of 5 produced weakened effect sizes.

Moreover, this study's definition of "fuzzy" family history was also conservative; only first degree relatives were eligible for this classification. Not including second degree relatives with diagnoses of uncertain validity or diagnosis that are frequently misdiagnosed may underestimate the rate of fuzzy bipolar; thus, resulting in weakened effect sizes as well. Also noteworthy, some research suggests that individuals with familial probands of bipolar

disorder are not five but ten times more likely to manifest the illness (Smoller & Finn, 2003). This study was consistently conservative in regards to quantifying familial risk status (e.g., DLR = 5 versus 10 for first degree relatives with bipolar). Because of the exploratory nature of this project (i.e., little research up to now has classified different categories of familial risk), researchers chose to take a conservative approach as it seemed like a prudent first step.

Another caveat is the possibility that family history information may not be entirely accurate or complete. The majority (88%) of family history information was gathered from participants' mothers' using a semi-structured interview, the MINI. It is anticipated that data collected using the MINI about a different person may reflect reporter bias to some extent. Related, because the majority of family history information was provided by participants' mothers, it is likely that in some cases family history information is incomplete. For example, some mothers may not remember or may have limited information about their own family history or that of their child's father. Further, only 4 second degree relatives out of 643 participants (and this is a set of several thousand relatives) received any bipolar diagnosis. The rate of detected bipolar is more than an order of magnitude lower than the rates found in epidemiological studies in the general United States population, suggesting that the research interview did not demonstrate pure diagnostic sensitivity to bipolar conditions. This finding is largely a result of the study design. Specifically, research interviewers most often only collected family history information for youths' relatives who were currently residing in the family household. Consequently, this design limitation may have weakened study results. In sum, a noteworthy limitation of the present study is the likely possibility that family history of bipolar illness was not captured for all relatives of youth participants and that some of the information provided from youths' primary caregivers may reflect reporter bias.

Moreover, the definition of fuzzy family history used for this study is potentially problematic. In theory, the fuzzy category is reserved for diagnoses of uncertain validity, or else a different diagnosis for which bipolar is often mistaken in a minority population (e.g., schizophrenia or conduct disorder in African-America populations) (DelBello et al., 2001). As a result of secondary analysis and pre-established study procedures, it was not possible to identify the race/ethnicity of youths' relatives. Because the sample was largely African-American and the literature indicates that African-Americans have a history of being misdiagnosed with psychotic disorders, all first degree relatives (regardless of race/ethnicity status) with a psychotic diagnosis resembling bipolar illness were assigned to the fuzzy category. This was the best solution to a limitation of secondary analysis of preexisting data. However, the challenges of vague family history and complex (or indeterminate) racial and ethnic composition will also be issues in clinical practice.

Additional limitations stem from the lack of prior research on threshold models in the MH field, specifically with bipolar disorder. Conceptually, threshold models are gaining popularity in the assessment of PBD (Youngstrom, Freeman, & Jenkins, 2009); however, no previous MH research has investigated specific benchmarks for the wait-test or test-treat thresholds. For example, what probability of bipolar is needed to cross the test-treatment threshold in order for clinicians to begin treatment (e.g., 85%, 90%, or 95%) (Kraemer, 1992; Straus et al., 2005; Swets, Dawes, & Monahan, 2000)? And, what role do patient preferences play in setting test-wait and the test-treatment thresholds? The present study selected relatively sensible thresholds; *85* was used for the test-treat threshold and *25* was used for the wait-treat threshold. Note that these specific thresholds influenced agreement between clinical and assessment methodologies about next clinical action (i.e., different thresholds

would theoretically result in a different kappa). Thresholds for the given project are somewhat arbitrary since no previous research is available to inform the numerical values of these thresholds. This clearly would be a productive area for future inquiry- differences in test-wait could be related to differences in treatment-seeking; test-treat could be related to treatment choice of adherence, etc. Nevertheless, this project is highly innovative and study findings, including findings regarding the relationship between clinician confidence and type of bipolar, stand to make a unique contribution to the literature, even if that of a preliminary nature.

Finally, starting base rate used for the present study represents a research challenge and broader clinical issue. In particular, the given study sample consisted of youth participants from two different clinical settings (a urban outpatient community mental health clinic and specialty research clinic) which complicated the process of determining the most appropriate starting base rate (i.e., there is no empirical support in the literature for base rates in blended settings). To address this challenge, investigators made an educated guess regarding the best starting base rate which took into account the separate base rates of the two settings as well as specific diagnostic information about the specific sample. For more information regarding how a starting base rate of 12% was determined for this study, see Methods section. This potential limitation is noteworthy as it not only influenced the actuarial estimates in the present study, but it also exemplifies a possible barrier for clinicians in the community. In particular, clinicians may struggle to establish accurate starting base rates if their setting does not align with those specified in the current literature. This is an important consideration that should be addressed in future research.

Future Directions

Despite the encouraging findings from the present study, substantial research is needed to elucidate the role of EB decision aids in psychological assessment. A question that might arise is whether clinicians on the front line are willing to use the nomogram in real-world practice. There is currently limited research investigating MH professionals' attitudes toward the nomogram. For example, do EB assessment methods *appeal* to clinicians as a tool for helping them make decisions in challenging diagnoses such as PBD? And, what do clinicians see as strengths and weaknesses of EB assessment tools in the assessment of PBD? Recent research suggests that a majority of MH professionals who received a brief nomogram training (< 30 minutes) endorsed using it in practice (Jenkins et al., 2008). Nevertheless, additional research is needed to more fully understand clinicians' perspectives. Qualitative methodologies may shed light on clinicians' attitudes and impressions that can inform future education and training of the nomogram; in turn, this knowledge can potentially increase the rate at which clinical practice absorbs EB assessment tools.

Along similar lines, more research is needed to understand potential barriers to implementing the nomogram in clinical practice. Specifically, before disseminating EBA tools, one must first consider barriers at the individual (i.e., provider and consumer) and agency (i.e., clinical setting) levels. Above and beyond gauging clinicians' enthusiasm, there is currently limited knowledge about barriers related to more general policy considerations (e.g., supervision, clinical training models, etc.). Increasing the knowledge base about these types of barriers can help increase the successful transfer of EB tools from laboratory to practice settings, a recurring impasse for the EBP movement in the field of clinical psychology.

Related to enhancing the transportability of EB assessment decision aids into clinical arenas, a number of different mechanisms exist for "packaging" actuarial assessment. Bayes' Theorem can be presented in a number of user-friendly ways. For example, rather than using paper copies of the nomogram, software packages could be designed that would allow the clinicians to simply input pieces of information (e.g., test score, family history) and all calculations would be performed behind the scenes. This approach would eliminate the need for clinicians to use the nomogram as a type of slide rule. Innovative technology combined with EB decision-making tools might greatly appeal to audiences that find statistics intimidating, or in settings that cannot facilitate nomogram training and supervision. Using technology to expedite the delivery of EB mental health services is a rapidly growing niche (Bucholz et al., 1991; Erdman et al., 1992; Finfgeld, 1999; Kobak et al., 1997) and may be an effective way for clinicians to adopt EB tools in practice. Nevertheless, more research is needed to examine the best way to present actuarial assessment methodologies to maximize their clinical utility.

Overall, research indicates that dissemination efforts are more successful when procedures are acceptable, feasible, and likely to permit adherence (Arndorfer, Allen, & Aliazireh, 1999). Better understanding clinician perspectives about actuarial assessment along with identifying potential practice barriers and the most attractive medium for Bayes' Theorem can enhance acceptance, feasibility, and adherence. Specifically, working with clinicians to maximize the clinical utility of actuarial methods is recommended for increasing rates of successful dissemination. Before attempting to create large scale change in current decision-making practice, researchers should first pilot EB tools with providers.

Further, two specific domains of the threshold model could benefit from additional

investigation. First, as previously indicated, no research to date has tested different test-wait and the test-treatment thresholds. Second, the current literature provides limited information about the role of patient preferences. For instance, it is unclear to what extent patient preferences influence the test-treatment threshold. Psychopharmacological agents used to treat bipolar disorder carry a number of potentially unpleasant side effects for youths, such as severe acne and weight gain. The perceived adversity of side effects will naturally vary on an individual basis. More work is needed to understand how the interaction of clinicians' recommendations and patient preferences map on to the threshold model. In other words, the scientific community needs to address the degree to which patient priorities impact EB assessment strategies and respective treatment recommendations.

Finally, current study analyses were based on a predominantly African-American sample. Given the exploratory nature of this project, it is recommended that study hypotheses be retested with other populations. It will be important to see if results are replicated with more diverse samples. Notwithstanding, it is highly encouraging that results generalized to this underserved group. EBA approaches demonstrate potential for making a substantial impact on helping reduce misdiagnosis among minorities.

Overall, this project provides an exploratory examination of clinical and actuarial approaches to the assessment of PBD. As such, the study represents the beginning of a program of research that investigates the role of EB tools in the diagnosis of complex mental disorders, such as bipolar illness. Altogether, findings from this study support the clinical utility of the nomogram and, more generally, actuarial assessment methodologies in clinical psychology research and practice.

Table 1

Glossary of Terms

Term	Aliases
Actuarial Bayesian Estimate	actuarial estimate Bayesian estimate Bayesian posterior probability likelihood of bipolar nomogram estimate post-test probability
Longitudinal Expert All Data (LEAD) Diagnosis	research diagnosis diagnosis based on clinical assessment gold standard clinical assessment
Clinician Confidence Rating	clinical assessment estimate clinician confidence confidence rating LEAD estimate LEAD rating

Table 2

Errors in Decision-Making

Heuristics, biases, & cognitive errors	Definition
Availability	Overestimating probability of a diagnosis when instances are relatively easy to recall.
Base Rate Neglect	Failing to adequately take into account prevalence of illness; representativeness exclusivity.
Confirmation	Selectively gathering and interpreting evidence that confirms a diagnosis and ignoring evidence that might disconfirm it.
Framing	Choosing riskier treatments when they are described in negative rather than positive terms.
Hindsight	Overestimating probability of a diagnosis when the correct diagnosis is already known.
Regret	Overestimating probability of a diagnosis with severe possible outcome because of anticipated regret if diagnosis were missed.
Representativeness	Overemphasizing evidence that strongly resembles a class of events; vulnerable to undervalue relevant base rates.
Unpacking principle	Providing more detail of an event increases its judged probability; also known as, effect of description.

Note. Source adapted from Elstein & Schwartz, 2002; Galanter & Patel, 2005.

Table 3

Participant Demographic and Diagnostic Characteristics

Characteristic	*n* (%)
Gender: Male	393 (61)
Ethnicity	
African-American	451 (70)
White	141 (22)
Hispanic	16 (3)
Other	32 (5)
Average # of Axis I Diagnoses	3.7 (*SD* = 1.69)
Reference Standard Positive[ab]	
Bipolar I	28 (4)
Bipolar II	17 (3)
Cyclothymic Disorder	51 (8)
Bipolar NOS	54 (8)
Reference Standard Negative[ab]	
Unipolar Depression (MDD, dysthymia, adjustment disorder)	180 (28)
ADHD or disruptive behavior without mood disorder	283 (44)
Residual (anxiety, PTSD, psychotic disorders, no Axis I)	51 (8)
Any ADHD	396 (62)

[a]Diagnoses were based on LEAD diagnoses. [b]Percentages add up to >100% due to comorbidity.

Table 4

DLRs Associated with Test Scores on the PGBI (28-item)

			Ages 5-10			
Range	Low	Mod. Low	Neutral	Mod. High	High	Very High
Score	> 11	11-20	21-30	31-42	43-50	51+
DLR	0.10	0.48	1.34	2.31	4.90	6.29
			Ages 11-17			
Range	Low	Mod. Low	Neutral	Mod. High	High	Very High
Score	> 9	9-15	16-24	25-39	40-48	49+
DLR	0.06	0.25	1.12	2.22	4.82	9.21

Note. From Youngstrom et al., 2004. Low = bottom 20% of sample; Mod. Low = moderately low, 21st to 40th percentile; Neutral = 41st to 60th percentile; Mod. High = moderately high, 61st to 80th percentile; High = 81st to 90th percentile; Very High = top 10%.

Table 5

Family Risk Status

Family history of bipolar illness	DLR	Frequency[a]	%
1st degree relative with BP-I	5	103	16
Second degree relative with BP-I or BP-II	2.5	4	< 1
"Fuzzy" family history of bipolar disorder	2	70	11
No family history of bipolar	1	466	73

Note. None of the youth participants who had a second degree relative with a bipolar diagnosis also had a first degree relative with bipolar. Therefore, it was not necessary to use a hierarchical model such that if a first and second degree relative have bipolar, a youth only gets counted in the top row (i.e., DLR = 5).

[a]$N = 643$.

Table 6

DLRs Assigned for MINI Diagnosis by Type of Relative

MINI diagnoses	Type of relative	
	First degree	Second degree
Broad bipolar phenotype		
BP I	5	2.5
BP II	2	2.5
Narrow bipolar phenotype		
Cyclothymic Disorder	2	
BP NOS	2	
Frequently misdiagnosed mental D/Os		
Any Schizophrenia	2	
Mood D/O NOS	2	
Mood D/O w/ psychotic features	2	
Brief Psychotic Disorder	2	
Psychotic Disorder NOS	2	
Depression with Atypical Features	2	
Delusional Disorder NOS	2	

Note. D/O = Disorder; NOS = Not Otherwise Specified; w/ = with.

Table 7

Variable Descriptives

Variable	*n*	Mean (*SD*)	Median	Skewness	Kurtosis
Clinician confidence ratings	627	15.28 (30.74)	0	1.67	.99
PGBI scores	636	21.28 (14.89)	19.00	.83	.37
Nomogram estimates	624	19.83 (20.74)	13.20	1.20	.57

Table 8

Agreement between Clinical and Actuarial Approaches (kappa = .21, $p < .0005$)

		LEAD Confidence Ratings			
		Mild Probability > 25% *Wait*	Med Probability 26-84% *Test*	High Probability >85% *Treat*	Total
Nomogram Estimates					
Mild Probability > 25% *Wait*	*n*	398	43	20	461
	% w/in actuarial	86%	9%	4%	100%
	% w/in clinical	82%	57%	43%	76%
Med Probability 26-84% *Test*	*n*	87	33	25	145
	% w/in actuarial	60%	23%	17%	100%
	% w/in clinical	18%	43%	53%	24%
High Probability >85% *Treat*	*n*	0	0	2	2
	% w/in actuarial	0%	0%	100%	100%
	% w/in clinical	0%	0%	4%	.3%
Total	N	485	76	47	608
	% w/in actuarial	80%	12%	8%	100%
	% within clinical	100%	100%	100%	100 %

Note. Med = Medium.

Table 9

Final Model: Regression Weights, Standard Error, and Significance for Regression Model

Predictor* (n = 136)	***B***	***SE***	***p***
Actuarial Bayesian Estimate	6.65	12.12	.584
LEAD Diagnostic Status			
BP-I Yes or No	31.29	11.80	.009
BP-II Yes or No	26.99	12.88	.038
Cyclothymic Disorder Yes or No	27.67	7.98	.001
Interaction Terms (Diagnosis * Actuarial Bayesian Estimate)			
BP-I	.28	24.43	.991
BP-II	8.79	33.02	.790
Cyclothymic Disorder	-2.60	19.45	.894
Constant	49.77	4.43	.000

**Note.* The significance of the interaction terms was the direct test of the hypothesis; none were significant, $p > .05$.

Figure 1. Agreement between Clinician Confidence and the Nomogram.

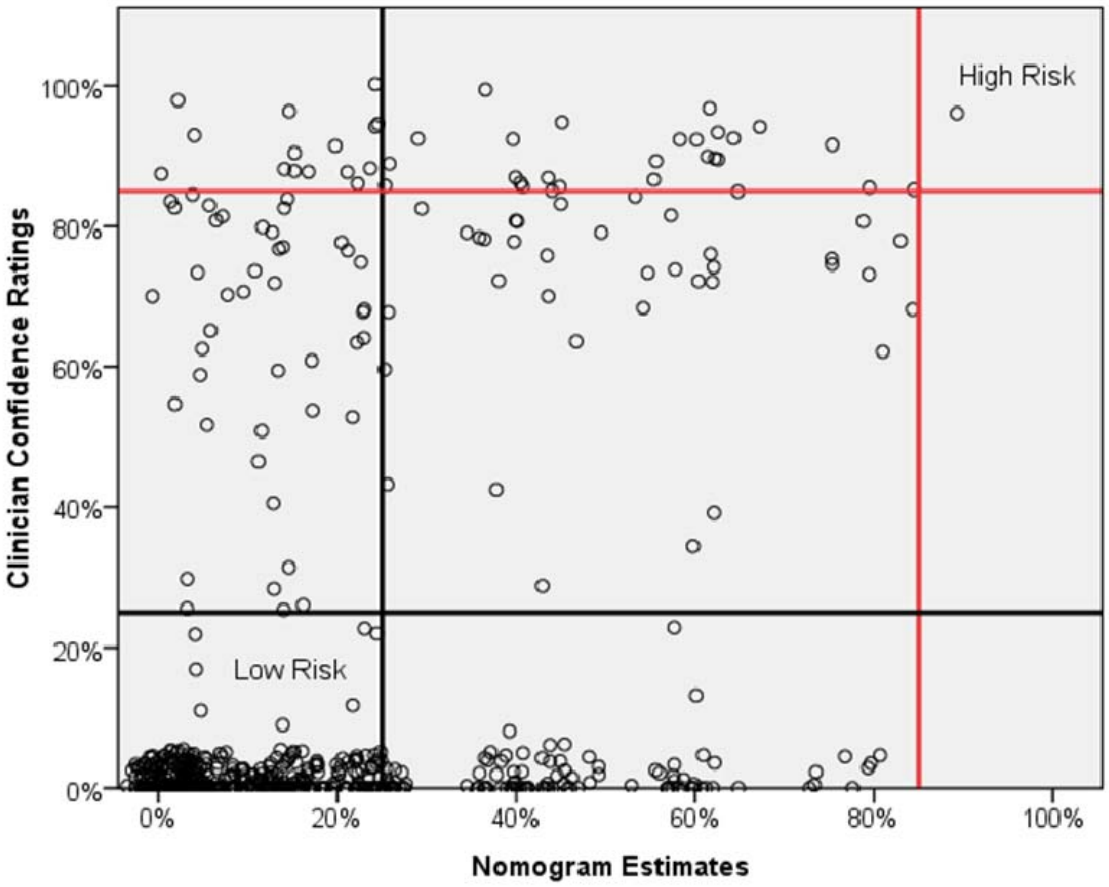


Figure 2. Clinician Confidence and Actuarial Estimates by Type of Bipolar.

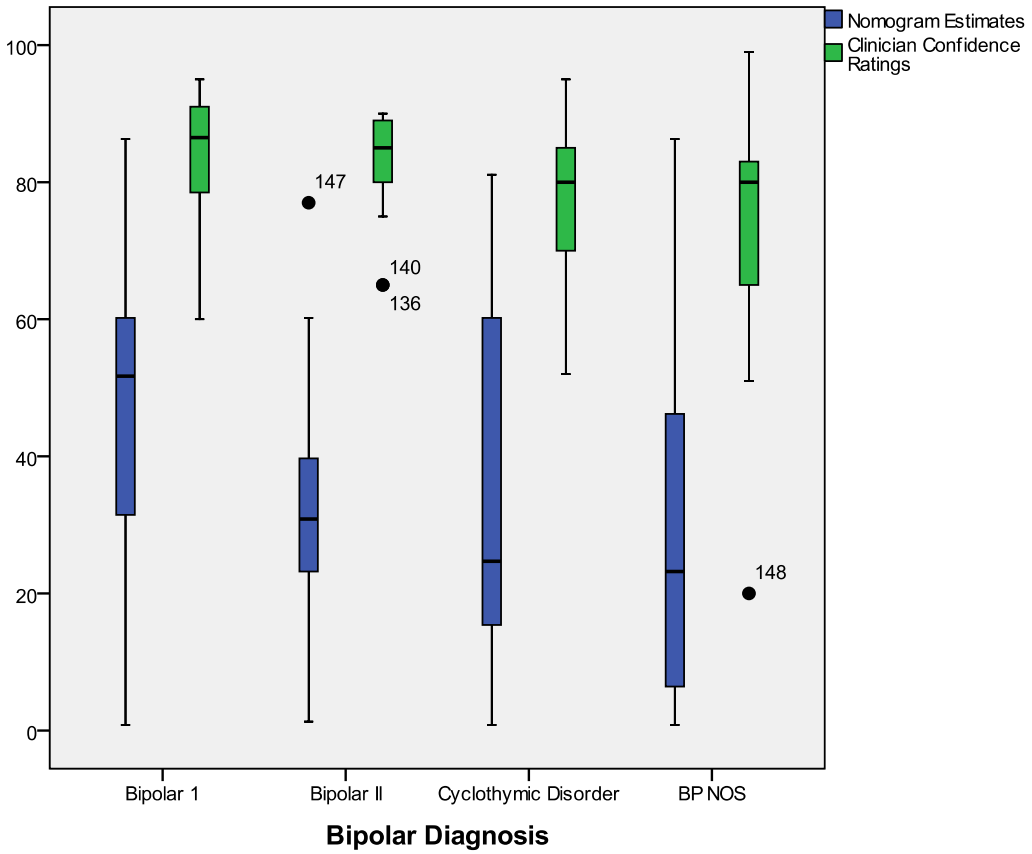


Figure 3. Diagnosis as Moderator of Agreement between Nomogram and Clinical Confidence.

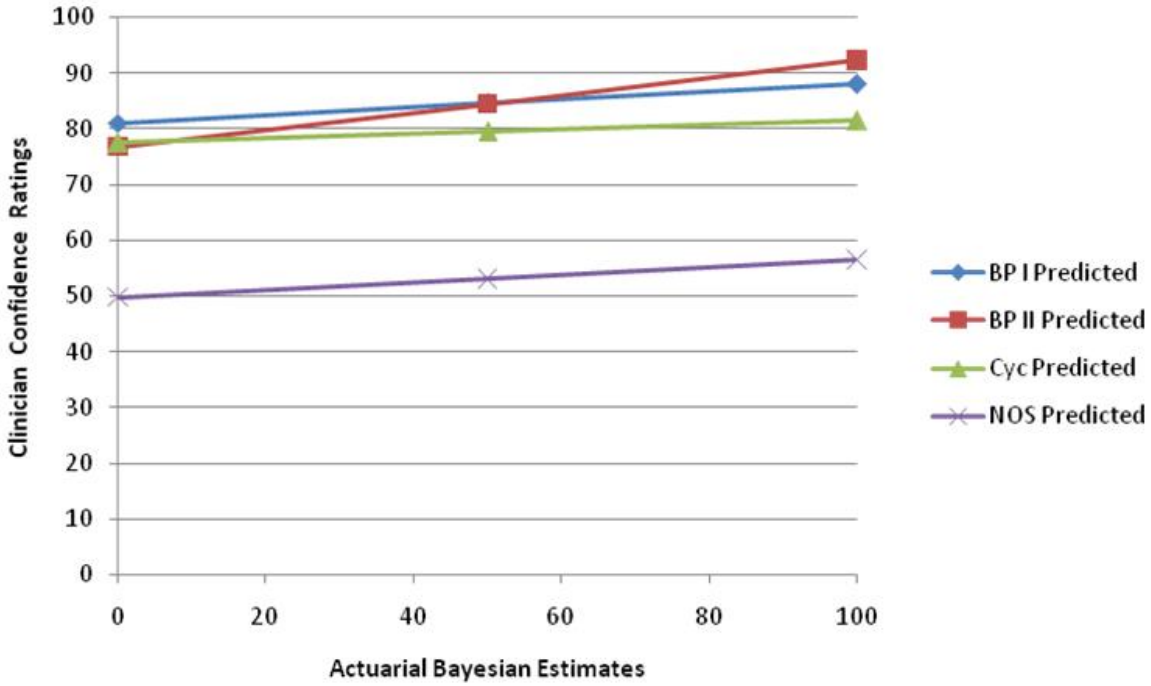
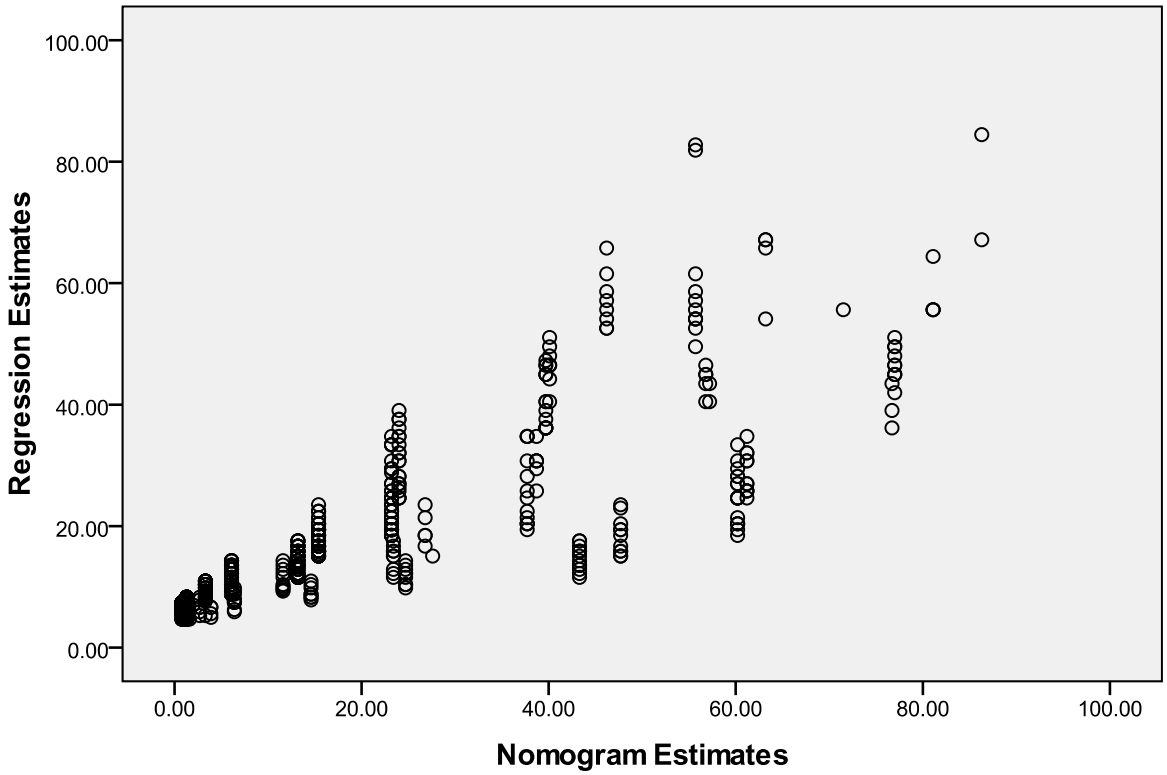
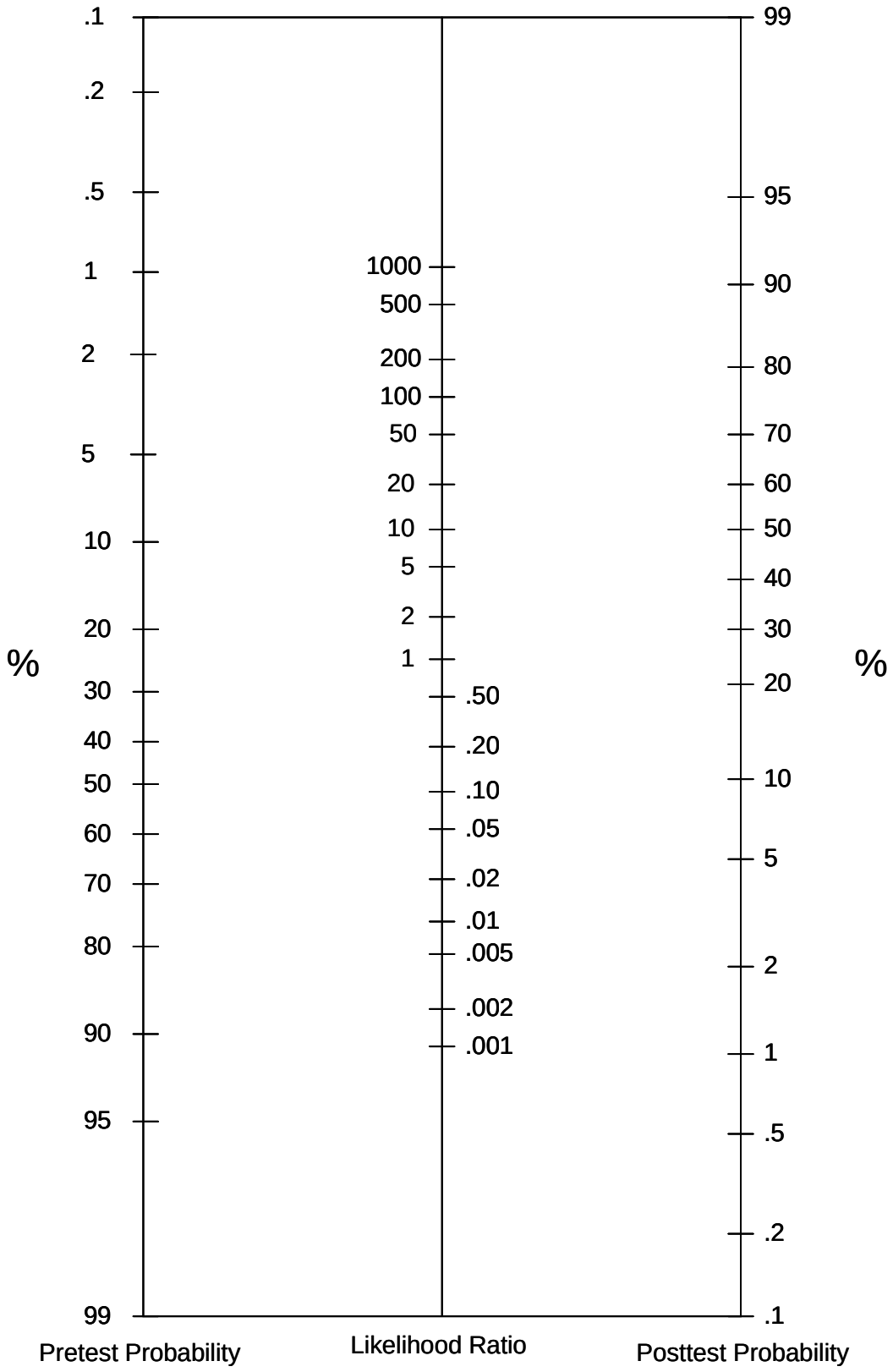


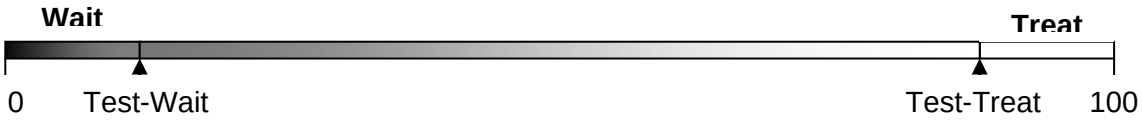
Figure 4. Correlation between Bayesian risk estimates and logistic regression estimates.



Appendix A. The Nomogram.



Appendix B. Threshold Model.



References

Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Teacher's Report Form and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991c). *Manual for the Youth Self Report form and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Allison, P. D. (2002). *Missing data*. Thousand Oaks: Sage.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Andreasen, N. C., Endicott, J., Spitzer, R. L., & Winokur, G. (1977). The family history method using diagnostic criteria. Reliability and validity. *Archives of General Psychiatry, 34*(10), 1229-1235.

Anthony, J. C., Folstein, M., Romanoski, A. J., Von Korff, M. R., Nestadt, G. R., Chahal, R., et al. (1985). Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis. Experience in eastern Baltimore. *Archives of General Psychiatry, 42*(7), 667-675.

Arndorfer, R. E., Allen, K. D., & Aliazireh, L. (1999). Behavioral health needs in pediatric medicine and the acceptability of behavioral solutions: Implications for behavioral psychologists. *Behavior Therapy, 30*(1), 137-148.

Axelson, D. A., Birmaher, B., Strober, M., Gill, M. K., Valeri, S., Chiappetta, L., et al. (2006). Phenomenology of children and adolescents with bipolar spectrum disorders. *Archives of General Psychiatry, 63*(10), 1139-1148.

Berument, S. K., Rutter, M., Lord, C., Pickles, A., & Bailey, A. (1999). Autism screening questionnaire: diagnostic validity. *British Journal of Psychiatry, 175*, 444-451.

Biederman, J., Faraone, S., Mick, E., Wozniak, J., Chen, L., Ouellette, C., et al. (1996). Attention-deficit hyperactivity disorder and juvenile mania: an overlooked comorbidity? *Journal of the American Academy of Child and Adolescent Psychiatry, 35*(8), 997-1008.

Blader, J. C., & Carlson, G. A. (2007). Increased rates of bipolar disorder diagnoses among U.S. child, adolescent, and adult inpatients, 1996-2004. *Biological Psychiatry, 62*(2), 107-114.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). Standards for Reporting of Diagnostic Accuracy steering group Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326*, 41-44.

Bowring, M. A., & Kovacs, M. (1992). Difficulties in diagnosing manic disorders among children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 31*(4), 611-614.

Bucholz, K. K., Robins, L. N., Shayka, J. J., Przybeck, T. R., Helzer, J. E., Goldring, E., et al. (1991). Performance of two forms of a computer psychiatric screening interview: version I of the DISSI. *Journal of Psychiatric Research, 25*(3), 117-129.

Carlson, G. A., & Youngstrom, E. A. (2003). Clinical implications of pervasive manic symptoms in children. *Biological Psychiatry, 53*, 1050-1058.

Cicchetti, D., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., et al. (2006). Rating scales, scales of measurement, issues of reliability: resolving some critical issues for clinicians and researchers. *Journal of Nervous and Mental Disease, 194*(8), 557-564.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*(2), 127-137.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum.

Croskerry, P. (2002). Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine, 9*(11), 1184-1204.

Danielson, C. K., Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Discriminative validity of the general behavior inventory using youth report. *Journal of Abnormal Child Psychology, 31*(1), 29-39.

Davidow, J., & Levinson, E. M. (1993). Heuristic principles and cognitive bias in decision making: Implications for assessment in school psychology. *Psychology in the Schools, 30*(4), 351-361.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

DelBello, M. P., Lopez-Larson, M. P., Soutullo, C. A., & Strakowski, S. M. (2001). Effects of race on psychiatric diagnosis of hospitalized adolescents: A retrospective chart review. *Journal of Child and Adolescent Psychopharmacology, 11*(1), 95-103.

Depue, R. A., Slater, J. F., Wolfstetter-Kausch, H., Klein, D. N., Goplerud, E., & Farr, D. A. (1981). A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: A conceptual framework and five validation studies. *Journal of Abnormal Psychology, 90*(5), 381-437.

Donner, A. (1998). Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine, 17*(10), 1157-1168.

Dubicka, B., Carlson, G. A., Vail, A., & Harrington, R. (2008). Prepubertal mania: diagnostic differences between US and UK clinicians. *European Child and Adolescent Psychiatry, 17*(3), 153-161.

Dunn, L. M., & Dunn, L. M. (1997). *Examiner's Manual for the Peabody Picture Vocabulary Test -- Third Edition*. Circle Pines, MN: American Guidance Service.

Dunner, D. L. (2003). Clinical consequences of under-recognized bipolar spectrum disorder. *Bipolar Disorders, 5*(6), 456-463.

Elstein, A. S., & Schwartz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *British Medical Journal, 324*(7339), 729-732.

Erdman, H. P., Klein, M. H., Greist, J. H., Skare, S. S., Husted, J. J., Robins, L. N., et al. (1992). A comparison of two computer-administered versions of the NIMH Diagnostic Interview Schedule. *Journal of Psychiatric Research, 26*(1), 85-95.

Ezpeleta, L., de la Osa, N., Domenech, J. M., Navarro, J. B., Losilla, J. M., & Judez, J. (1997). Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents-DICA-R-in an outpatient sample. *Journal of Child Psychology and Psychiatry, 38*(4), 431-440.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G^* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175.

Findling, R. L., Gracious, B. L., McNamara, N. K., Youngstrom, E. A., Demeter, C., & Calabrese, J. R. (2001). Rapid, continuous cycling and psychiatric co-morbidity in pediatric bipolar I disorder. *Bipolar Disorders, 3*, 202-210.

Finfgeld, D. L. (1999). Computer-based mental health assessments. Panaceas, pariahs, or partners in research and practice? *Computers in Nursing, 17*(5), 215-220.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd ed.). New York: Wiley.

Fleiss, J. L. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). New York: Wiley.

Fletcher, J. M., Francis, D. J., Morris, R. D., & Lyon, G. R. (2005). Evidence-based assessment of learning disabilities in children and adolescents. *Journal of Clinical Child Adolescent Psychology, 34*(3), 506-522.

Gaissmaier, W., & Gigerenzer, G. (2008). Statistical illiteracy undermines informed shared decision making. *German Journal for Evidence and Quality in Health Care, 102*(7), 411-413.

Galanter, C. A., & Patel, V. L. (2005). Medical decision making: a selective review for child psychiatrists and psychologists. *Journal of Child Psychology and Psychiatry, 46*(7), 675-689.

Geller, B., Craney, J. L., Bolhofner, K., DelBello, M. P., Williams, M., & Zimerman, B. (2001). One-year recovery and relapse rates of children with a prepubertal and early adolescent bipolar disorder phenotype. *The American Journal of Psychiatry, 158*(2), 303-305.

Geller, B., Craney, J. L., Bolhofner, K., Nickelsburg, M. J., Williams, M., & Zimerman, B. (2002). Two-Year Prospective Follow-Up of Children With a Prepubertal and Early Adolescent Bipolar Disorder Phenotype. *American Journal of Psychiatry, 159*, 927-933.

Geller, B., Tillman, R., Craney, J. L., & Bolhofner, K. (2004). Four-year prospective outcome and natural history of mania in children with a prepubertal and early adolescent bipolar disorder phenotype. *Archives of General Psychiatry, 61*(5), 459-467.

Geller, B., Zimerman, B., Williams, M., Bolhofner, K., Craney, J. L., DelBello, M. P., et al. (2001). Reliability of the Washington University in St. Louis Kiddie Schedule for Affective Disorders and Schizophrenia (WASH-U-KSADS) mania and rapid cycling sections. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(4), 450-455.

Ghaemi, N., Sachs, G. S., & Goodwin, F. K. (2000). What is to be done? Controversies in the diagnosis and treatment of manic-depressive illness. *World Journal of Biological Psychiatry, 1*(2), 65-74.

Ghaemi, S. N., Bauer, M., Cassidy, F., Malhi, G. S., Mitchell, P., Phelps, J., et al. (2008). Diagnostic guidelines for bipolar disorder: a summary of the International Society for Bipolar Disorders Diagnostic Guidelines Task Force Report. *Bipolar Disorders, 10*(1 Pt 2), 117-128.

Ghaemi, S. N., Sachs, G. S., Chiou, A. M., Pandurangi, A. K., & Goodwin, F. K. (1999). Is bipolar disorder still underdiagnosed? Are antidepressants overutilized? *Journal of Affective Disorders, 52*(1), 135-144.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 513-525.

Gracious, B. L., Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2002). Discriminative validity of a parent version of the Young Mania Rating Scale. *Journal of the American Academy of Child and Adolescent Psychiatry, 41*, 1350-1359.

Gray, G. E. (2005). *Consise Guide to Evidence-Based Psychiatry*. Washington, D.C.: American Psychiatric Publishing, Inc.

Guyatt, G. H., & Rennie, D. (Eds.). (2002). *Users' guides to the medical literature*. Chicago: AMA Press.

Helzer, J. E., Robins, L. N., McEvoy, L. T., Spitznagel, E. L., Stoltzman, R. K., Farmer, A., et al. (1985). A comparison of clinical and diagnostic interview schedule diagnoses. Physician reexamination of lay-interviewed cases in the general population. *Arch Gen Psychiatry, 42*(7), 657-666.

Hirschfeld, R. M., Calabrese, J. R., Weissman, M. M., Reed, M., Davies, M. A., Frye, M. A., et al. (2003). Screening for bipolar disorder in the community. *Journal of Clinical Psychiatry, 64*(1), 53-59.

Hirschfeld, R. M., Lewis, L., & Vornik, L. A. (2003). Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry, 64*(2), 161-174.

Hodgins, S., Faucher, B., Zarac, A., & Ellenbogen, M. (2002). Children of parents with bipolar disorder. A population at high risk for major affective disorders. *Child and Adolescent Psychiatric Clinics of North America, 11*(3), 533-553.

Hunink, M., Glasziou, P., Siegel, J., Elstein, A., Weeks, J., Pliskin, J., et al. (2001). Decision Making in Health and Medicine: Integrating Evidence and Values. New York: Cambridge University Press.

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994a). Users' guides to the medical literature: III. How to use an article about a diagnostic test: A. Are the results of the study valid? *Journal of the American Medical Association, 271*(5), 389-391.

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994b). Users' guides to the medical literature: III. How to use an article about a diagnostic test: B: What are the results and will they help me in caring for my patients? *Journal of the American Medical Association, 271*(9), 703-707.

Jenkins, M. M., Youngstrom, E. A., Youngstrom, J. K., & Algorta, G. P. (2008). *How the Nomogram Improves Interpretation of Assessment Information by Clinicians in the Community*. Paper presented at the Annual Meeting for the Association for Behavioral and Cognitive Therapy (ABCT).

Joseph, M. F., Youngstrom, E. A., & Soares, J. C. (2009). Antidepressant-coincident mania in children and adolescents treated with selective serotonin reuptake inhibitors. *Future Neurology., 4*(1), 87-102.

Judd, L. L., & Akiskal, H. S. (2003). The prevalence and disability of bipolar spectrum disorders in the US population: re-analysis of the ECA database taking into account subthreshold cases. *Journal of Affective Disorders, 73*(1-2), 123-131.

Kadri, N., Agoub, M., El Gnaoui, S., Alami, K. M., Hergueta, T., & Moussaoui, D. (2005). Moroccan colloquial Arabic version of the Mini International Neuropsychiatric Interview (MINI): Qualitative and quantitative validation. *European Psychiatry, 20*(2), 193-195.

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., et al. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*(7), 980-988.

Kessler, R. C. (1999). Comorbidity of unipolar and bipolar depression with other psychiatric disorders in a general population survey. In M. Tohen (Ed.), *Comorbidity in Affective Disorders* (pp. 1-25). New York: Marcel Dekker, Inc.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 617-627.

Kirov, G., & Murray, R. M. (1999). Ethnic differences in the presentation of bipolar affective disorder. *European Psychiatry, 14*(4), 199-204.

Klassen, A. F., Miller, A., & Fine, S. (2006). Agreement between parent and child report of quality of life in children with attention-deficit/hyperactivity disorder. *Child Care Health and Development, 32*(4), 397-406.

Kluger, J., & Song, S. (2002). Young and bipolar. *Time*, *August 19*, 39– 47, 51.

Kobak, K. A., Taylor, L. H., Dottl, S. L., Greist, J. H., Jefferson, J. W., Burroughs, D., et al. (1997). Computerized screening for psychiatric disorders in an outpatient community mental health clinic. *Psychiatric Services, 48*(8), 1048-1057.

Kowatch, R. A., Youngstrom, E. A., Danielyan, A., & Findling, R. L. (2005). Review and meta-analysis of the phenomenology and clinical characteristics of mania in children and adolescents. *Bipolar Disorders, 7*, 483-496.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage Publications.

Kraepelin, E. (1921). *Manic-depressive insanity and paranoia*. Edinburgh: Livingstone.

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., et al. (2004). Achieving competency in psychological assessment: directions for education and training. *Journal of Clinical Psychology 60*(7), 725-739.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lau AYS, Coiera EW. (2007) How do clinicians search for and access biomedical literature to answer clinical questions? MEDINFO 2007; Brisbane, Australia.

Lecrubier, Y., Sheehan, D. V., Weiller, E., Amorim, P., Bonora, I., Sheehan, K. H., et al. (1997). The Mini International Neuropsychiatric Interview (MINI): A short diagnostic structured interview: Reliability and validity according to the CIDI. *European Psychiatry, 12*(5), 224-231.

Leibenluft, E., Charney, D. S., Towbin, K. E., Bhangoo, R. K., & Pine, D. S. (2003). Defining clinical phenotypes of juvenile mania. *The American Journal of Psychiatry, 160*, 430-437.

Lewczyk, C. M., Garland, A. F., Hurlburt, M. S., Gearity, J., & Hough, R. L. (2003). Comparing DISC-IV and clinician diagnoses among youths receiving public mental health services. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*(3), 349-356.

Lewinsohn, P. M., Klein, D. N., & Seeley, J. R. (1995). Bipolar disorders in a community sample of older adolescents: Prevalence, phenomenology, comorbidity, and course. *Journal of the American Academy of Child and Adolescent Psychiatry, 34*(4), 454-463.

Loranger, A., Susman, V., Oldham, J., & Russakoff, L. (1987). The Personality Disorder Examination (PDE): A preliminary report. *Journal of Personality Disorders, 1*, 1-13.

Lueger, R. J. (2002). Practice-informed research and research-informed psychotherapy. *Journal of Clinical Psychology, 58*(10), 1265-1276.

Marchand, W. R., Wirth, L., & Simon, C. (2006). Delayed diagnosis of pediatric bipolar disorder in a community mental health setting. *Journal of Psychiatric Practice, 12*(2), 128-133.

Mash, E. J., & Hunsley, J. (2005). Evidence-Based Assessment of Child and Adolescent Disorders: Issues and Challenges. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 362-379.

Mass, J. (2003). Evidence-based assessment of children with behavioral and emotional disorders. *Report of Emotional and Behavioral Disorders in Youth, 3*, 31-34.

McClellan, J. M., Werry, J. S., & Ham, M. (1993). A follow-up study of early onset psychosis: comparison between outcome diagnoses of schizophrenia, mood disorders, and personality disorders. *Journal of Autism and Developmental Disorders, 23*(2), 243-262.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Merikangas, K. R, Herrell, R., Swendsen, J., Rossler, W., Ajdacic-Gross, V., & Angst, J. (2008). Specificity of bipolar spectrum conditions in the comorbidity of mood and substance use disorders: results from the Zurich cohort study. *Archives of General Psychiatry, 65*(1), 47-52.

Miller, C. J., Klugman, J., Berv, D. A., Rosenquist, K. J., & Ghaemi, S. N. (2004). Sensitivity and specificity of the Mood Disorder Questionnaire for detecting bipolar disorder. *Journal of Affective Disorders, 81*(2), 167-171.

Miller, P. R. (2001). Inpatient diagnostic assessments: 2. Interrater reliability and outcomes of structured vs. unstructured interviews. *Psychiatry Research, 105*(3), 265-271.

Miller, P. R. (2002). Inpatient diagnostic assessments: 3. Causes and effects of diagnostic imprecision. *Psychiatry Research, 111*(2-3), 191-197.

Miller, P. R., Dasher, R., Collins, R., Griffiths, P., & Brown, F. (2001). Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. *Psychiatry Research, 105*(3), 255-264.

Moreno, C., Laje, G., Blanco, C., Jiang, H., Schmidt, A. B., & Olfson, M. (2007). National trends in the outpatient diagnosis and treatment of bipolar disorder in youth. *Archives of General Psychiatry, 64*(9), 1032-1039.

Mukherjee, S., Shukla, S., Woodle, J., Rosen, A. M., & Olarte, S. (1983). Misdiagnosis of schizophrenia in bipolar patients: a multiethnic comparison. *American Journal of Psychiatry, 140*(12), 1571-1574.

Neighbors, H. W., Trierweiler, S. J., Ford, B. C., & Muroff, J. R. (2003). Racial Differences in DSM Diagnosis Using a Semi-Structured Instrument: The Importance of Clinical Judgment in the Diagnosis of African Americans. *Journal of Health and Social Behavior, 44*(3), 237-256.

Neighbors, H. W., Trierweiler, S. J., Munday, C., Thompson, E. E., Jackson, J. S., Binion, V. J., et al. (1999). Psychiatric diagnosis of African Americans: diagnostic divergence in clinician-structured and semistructured interviewing conditions. *Journal of the National Medical Association, 91*(11), 601-612.

Neisworth, J. T., & Bagnato, S. J. (2004). The Mismeasure of Young Children: The Authentic Assessment Alternative. *Infants and Young Children, 17*(3), 198-212.

Nottelmann, E., Biederman, J., Birmaher, B., Carlson, G. A., Chang, K. D., Fenton, W. S., et al. (2001). National Institute of Mental Health research roundtable on prepubertal bipolar disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(8), 871-878.

Olfson, M., Marcus, S. C., & Shaffer, D. (2006). Antidepressant drug therapy and suicide in severely depressed children and adults: A case-control study. *Archives of General Psychiatry, 63*(8), 865-872.

Otsubo, T., Tanaka, K., Koda, R., Shinoda, J., Sano, N., Tanaka, S., et al. (2005). Reliability and validity of Japanese version of the Mini-International Neuropsychiatric Interview. *Psychiatry and Clinical Neurosciences, 59*(5), 517-526.

Papolos, D. F., & Papolos, J. (1999). *The bipolar child: The definitive and reassuring guide to childhood's most misunderstood disorder*. New York: Broadway Books.

Pappadopulos, E., Jensen, P. S., Schur, S. B., MacIntyre, J. C., 2nd, Ketner, S., Van Orden, K., et al. (2002). "Real world" atypical antipsychotic prescribing practices in public child and adolescent inpatient settings. *Schizophrenia Bulletin, 28*(1), 111-121.

Paulos, J. A. (1988). *Innumeracy: Mathemathical illiteracy and its consequences.* New York: Vintage Books.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York: Wiley.

Peters, L., & Andrews, G. (1995). Procedural validity of the computerized version of the Composite International Diagnostic Interview (CIDI-Auto) in the anxiety disorders. *Psychological Medicine, 25*(6), 1269-1280.

Peterson, D. R. (2004). Science, Scientism, and Professional Responsibility. *Clinical Psychology: Science and Practice, 11*(2), 196-210.

Phelps, J., Angst, J., Katzow, J., & Sadler, J. (2008). Validity and utility of bipolar spectrum models. *Bipolar Disorders, 10*, 179-193.

Piacentini, J., Shaffer, D., Fisher, P., Schwab-Stone, M., Davies, M., & Gioia, P. (1993). The Diagnostic Interview Schedule for Children-Revised Version (DISC-R): III. Concurrent criterion validity. *Journal of American Academy of Child and Adolescent Psychiatry, 32*(3), 658-665.

Pilkonis, P. A., & Frank, E. (1988). Personality pathology in recurrent depression: nature, prevalence, and relationship to treatment response. *American Journal of Psychiatry, 145*(4), 435-441.

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*(1), 46-54.

Poznanski, E. O., Miller, E., Salguero, C., & Kelsh, R. C. (1984). Preliminary studies of the reliability and validity of the Children's Depression Rating Scale. *Journal of the American Academy of Child Psychiatry, 23*(2), 191-197.

Reimherr, J. P., & McClellan, J. M. (2004). Diagnostic challenges in children and adolescents with psychotic disorders. *Journal of Clinical Psychiatry, 65 Suppl 6*, 5-11.

Sackett, D. L., Haynes, R. B., & Guyatt, G. H. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine* (2nd ed.). Boston: Little Brown.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM* (2nd ed.). New York: Churchill Livingstone.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*(3), 380-400.

Shea, M. T., Glass, D. R., Pilkonis, P. A., & Watkins, J. (1987). Frequency and implications of personality disorders in a sample of depressed outpatients. *Journal of Personality Disorders, 1*(1), 27-42.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., et al. (1998). The Mini-International Neuropsychiatric Interview(M. I. N. I.): The development and validation of a Structured Diagnostic Psychiatric Interview for DSM-IV and ICD-10. *The Journal of clinical psychiatry. Supplement, 59*(20), 22-33.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Janavs, J., Weiller, E., Keskiner, A., et al. (1997). The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry, 12*(5), 232-241.

Smarty, S., & Findling, R. L. (2007). Psychopharmacology of pediatric bipolar disorder: a review. *Psychopharmacology, 191*(1), 39-54.

Smoller, J. W., & Finn, C. T. (2003). Family, twin, and adoption studies of bipolar disorder. *American Journal of Medical Genetics,123C*(1), 48-58.

Soutullo, C. A., Chang, K. D., Diez-Suarez, A., Figueroa-Quintana, A., Escamilla-Canales, I., Rapado-Castro, M., et al. (2005). Bipolar disorder in children and adolescents: international perspective on epidemiology and phenomenology. *Bipolar Disorders, 7*, 497-506.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24*(5), 399-411.

Sprock, J. (1988). Classification of schizoaffective disorder. *Comprehensive Psychiatry, 29*(1), 55-71.

Strakowski, S. M., Keck, P. E., Jr., Arnold, L. M., Collins, J., Wilson, R. M., Fleck, D. E., et al. (2003). Ethnicity and diagnosis in patients with affective disorders. *The Journal of Clinical Psychiatry, 64*(7), 747-754.

Strakowski, S. M., McElroy, S. L., Keck, P. E., Jr., & West, S. A. (1996). Racial influence on diagnosis in psychotic mania. *Journal of Affective Disorders, 39*(2), 157-162.

Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). New York: Churchill Livingstone.

Strober, M., Schmidt-Lackner, S., Freeman, R., Bower, S., & al, e. (1995). Recovery and relapse in adolescents with bipolar affective illness: A five-year naturalistic, prospective follow-up. *Journal of the American Academy of Child and Adolescent Psychiatry, 34*(6), 724-731.

Swets, J. A., Dawes, R. M., & Monahan, J. L. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26.

Todd, P. M., & Gigerenzer, G. (2007). Environments That Make Us Smart: Ecological Rationality. *Current Directions in Psychological Science, 16*(3), 167-171.

Tondo, L., Isacsson, G., & Baldessarini, R. J. (2003). Suicidal behaviour in bipolar disorder: Risk and prevention. *CNS Drugs, 17*(7), 491-511.

Tsuchiya, K. J., Byrne, M., & Mortensen, P. B. (2003). Risk factors in relation to an emergence of bipolar disorder: A systematic review. *Bipolar Disorders, 5*(4), 231-242.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Vitiello, B., Malone, R., Buschle, P. R., Delaney, M. A., & Behar, D. (1990). Reliability of DSM-III diagnoses of hospitalized children. *Hospital & Community Psychiatry, 41*(1), 63-67.

Vitiello, B., & Stoff, D. M. (1997). Subtypes of aggression and their relevance to child psychiatry. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*(3), 307-315.

Weissman, M. M., Wickramaratne, P., Adams, P., Wolk, S., Verdeli, H., & Olfson, M. (2000). Brief screening for family psychiatric history: the family history screen. *Archives of General Psychiatry, 57*(7), 675-682.

Weissman, M. M., Wickramaratne, P., Warner, V., John, K., Prusoff, B. A., Merikangas, K. R., et al. (1987). Assessing psychiatric disorders in children: Discrepancies between mothers' and children's reports. *Archives of General Psychiatry, 44*(8), 747-753.

Weller, E. B., Danielyan, A. K., & Weller, R. A. (2004). Somatic treatment of bipolar disorder in children and adolescents. *Psychiatric Clinics of North America, 27*(1), 155-178.

Wilens, T. E., Biederman, J., Kwon, A., Chase, R., Greenberg, L., Mick, E., et al. (2003). A systematic chart review of the nature of psychiatric adverse events in children and adolescents treated with selective serotonin reuptake inhibitors. *Journal of Child and Adolescent Psychopharmacology, 13*(2), 143-152.

Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: Reliability, validity, and sensitivity. *British Journal of Psychiatry, 133*, 429-435.

Youngstrom, E. A. (2007). Pediatric Bipolar Disorder. In E. J. Mash & R. A. Barkley (Eds.), *Assessment of Childhood Disorders* (pp. 253-304). New York: Guilford Press.

Youngstrom, E. A. (2008). Evidence-based strategies for the assessment of developmental psychopathology: Measuring prediction, prescription, and process. *Developmental psychopathology*, 34–77.

Youngstrom, E. A., Birmaher, B., & Findling, R. L. (2008). Pediatric bipolar disorder: validity, phenomenology, and recommendations for diagnosis. *Bipolar Disorders, 10*(1 Pt 2), 194-214.

Youngstrom, E. A., & Duax, J. (2005). Evidence based assessment of pediatric bipolar disorder, part 1: Base rate and family history. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*, 712-717.

Youngstrom, E. A., Findling, R. L., Calabrese, J. R., Gracious, B. L., Demeter, C., DelPorto Bedoya, D., et al. (2004). Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *Journal of the American Academy of Child and Adolescent Psychiatry, 43*, 847-858.

Youngstrom, E. A., Findling, R. L., Danielson, C. K., & Calabrese, J. R. (2001). Discriminative validity of parent report of hypomanic and depressive symptoms on the General Behavior Inventory. *Psychological Assessment, 13*(2), 267-276.

Youngstrom, E. A., Findling, R. L., Youngstrom, J. K., & Calabrese, J. R. (2005). Toward an evidence-based assessment of pediatric bipolar disorder. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 433-448.

Youngstrom, E. A., Freeman, A. J., & Jenkins, M. M. (2009). The assessment of children and adolescents with bipolar disorder. *Child and Adolescent Psychiatric Clinics of North America, 18*(2), 353-390, viii-ix.

Youngstrom, E. A., & Kogos Youngstrom, J. (2005). Evidence based assessment of pediatric bipolar disorder, part 2: Incorporating information from behavior checklists. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*, 823-828.

Youngstrom, E. A., Meyers, O., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006). Diagnostic and measurement issues in the assessment of pediatric bipolar disorder: Implications for understanding mood disorder across the life cycle. *Development and Psychopathology, 18*, 989-1021.

Youngstrom, E. A., Meyers, O. I., Demeter, C., Kogos Youngstrom, J., Morello, L., Piiparinen, R., et al. (2005). Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders, 7*(Special Issue: Pediatric Bipolar Disorder), 507-517.

Youngstrom, E. A., Youngstrom, J. K., & Starr, M. (2005). Bipolar Diagnoses in Community Mental Health: Achenbach CBCL Profiles and Patterns of Comorbidity. *Biological Psychiatry, 58*, 569-575.

Zarin, D. A., & Earls, F. (1993). Diagnostic decision making in psychiatry. *The American Journal of Psychiatry, 150*, 197-206.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition, 98*(3), 287-308.